

All together now

The decision to site the fusion experiment ITER in France left relatively little bad blood between the international partners, who must now rally behind the project.

So ITER — the international fusion-power experiment whose faltering progress sometimes seems to echo that of fusion research itself — may finally be built, after all. The countries involved have agreed in principle that construction should begin in France next year.

No one will question the technical capabilities of the hosts: France has an awesome tradition in nuclear technology, a strong will to make the project happen, and firm backing from the rest of the European Union (EU). A few reservations remain about the technical approach taken in ITER's design (see page 318), but most fusion researchers are delighted that the project looks set to proceed.

Some important details need to be resolved before that happens, however. Funding for the project is still to be lined up by most of the partners, for example, and much of it is likely to come at the expense of other, existing fusion research projects. Broadly speaking, the EU is supposed to pay half of the construction costs, while the other five partners — Japan, the United States, Russia, China and South Korea — pay 10% each. Additionally the nature of secondary facilities, to be built in Japan, has yet to be determined.

In several member countries, ITER will clash with domestic research priorities. In the United States, for example, the Bush administration has already tried once to shoehorn the \$50 million annual cost of ITER construction into the existing \$230 million budget for magnetic fusion research. Sherwood Boehlert, chairman of the House science committee, has rightly warned that this won't wash. If the administration is sincere in its support for a project that the Department of Energy has selected as its top-priority facility, it will allocate extra funds for ITER in its 2006 budget proposal, which comes out in February.

Some in Congress are bound to question support for any international project — especially one in France. But if the US scientific community unites behind the project, then a desire to treat the rest

of the world with less than total contempt will prevail in Congress, as happened when a similar amount of money was successfully appropriated over many years for the US contribution to the Large Hadron Collider in Switzerland.

The second issue concerns the types of supporting facilities to be built and their funding. Such projects are likely to include a materials testing centre, a computing centre for data analysis, and an upgrade for Japan's JT-60 fusion experiment. In return for Japan's agreement to drop its bid to build ITER itself, the EU will support Japan in its bid to lead these projects. Japan is best qualified and best equipped to do this. It should step up as a true leader, as it failed to do when pursuing ITER's construction. Other countries should support Japan in this role.

Among these projects, the most expensive — and the most valuable from the point of view of international progress towards fusion energy — would be a neutron source for use in materials testing. The crystalline structure of stainless steels and other metal alloys that might be used in working fusion reactors is expected to deteriorate rapidly under neutron bombardment. In the absence of a test facility that can supply a suitable neutron flux, no one has been able to search for metals or ceramics that might survive this bombardment for the lifetime of a working fusion reactor. Such a facility is needed, alongside ITER, to take magnetic fusion forward.

But progress has been slower than fusion advocates would like. To be fair, there has been a chicken-and-egg aspect to this: investment has been withdrawn from magnetic fusion research when it was most badly needed. Naysayers joke that fusion power has always been 50 years in the future — and always will be. Their scepticism needs to be balanced against the unique and almost boundless potential of fusion, should it be harnessed. Anyone who doubts this potential should try getting up early one morning to watch the sunrise. ■

Socialism in one country

Cuba's scientific community has made substantial progress in addressing social problems.

Despite a floundering economy, restrictions on free speech and the incessant hostility of its powerful neighbour to the north, Cuba has developed a considerable research capability — perhaps more so than any other developing country outside southeast Asia. Whatever one thinks of its leader, Fidel Castro, it is worth asking how Cuba did it, and what lessons other countries might draw from it.

When Castro came to power in 1959, Cuba had almost no scientific

infrastructure. Now it boasts a biotechnology industry that has produced effective drugs and vaccines of its own, a large and fairly influential scientific work-force, and a fledgling pharmaceutical industry with its sights set on export markets. The agricultural sector, in which small farmers benefit from partnerships with agricultural researchers, is also quite successful (see page 322).

Some of the reasons for Cuba's success are straightforward. The government has invested heavily in elementary and secondary education, and has attained developed-world standards of literacy and numeracy in its population. After university, large numbers of young scientists are sent abroad for training — once to its Communist allies, more recently to Europe and Latin America — and Cuba ensures, by fair means or foul, that they return home afterwards to work.

But one aspect of Cuba's scientific success is often overlooked. At



various times, other Latin American nations such as Venezuela and Argentina have sought to build up science and technology by supporting a mixture of pure and applied research, a model similar to that established in wealthier countries. Cuba took a different approach: research there is ruthlessly applied.

Cuba's state-sponsored science is structured like a corporate research laboratory, except that its output consists of social outcomes, rather than commercial products. If a project looks likely to earn foreign currency or meet the government's social objectives, it is backed to the hilt. Cuba's scientists have no funds for basic research, but they largely back their government's approach, in part because they have seen how it transformed health services in their country.

But the approach has many drawbacks. One concerns the constraints that it places on the movement of researchers. Castro's government maintains strict control on the movement of its citizens. Scientists fare better than most, and are frequently allowed to attend conferences or spend time working in foreign laboratories. Yet if they stay away for longer than permitted, they lose the right to return freely. This draconian approach to dealing with the threat of a brain drain is in breach of the Universal Declaration of Human Rights, adopted by the United Nations in 1948. Restrictions on free political expression in Cuba are also inconsistent with the declaration.

It is questionable, in any case, whether such restrictions serve any useful purpose for Cuba's government, given the obvious commit-

ment of the scientists in question to their country's future. Just as questionable is the purpose served by the continuing US trade embargo on Cuba, which continues to isolate scientists and others on the island from their colleagues in the United States, including a large group of Cuban origin.

The embargo damages Cuban science and scientific collaboration in various ways. A Cuban proposal for dengue research, for example, won a \$700,000 award from the Bill & Melinda Gates Foundation after international peer review. But the award has been held up for a year, lest the illustrious Microsoft founder, his wife and their fellow trustees be dragged off to the penitentiary for breaching the embargo.

Nature has consistently opposed scientific embargos, and strongly believes in research collaboration as a means of building bridges between nations that lack normal diplomatic relations. But there is a more specific issue here. When Castro dies, Cuba faces a period of volatility that could endanger key national assets, such as its science. In preparation for that day, both Havana and Washington should be acting now to wind down such cold-war artefacts as Cuba's travel restrictions and the US trade embargo. ■

"Cuba's science is structured like a corporate research lab, except that its output consists of social outcomes not commercial products."

Agency under siege

Conflicts-of-interest at the US National Institutes of Health justify the agency's ethics crackdown.

The latest information to emerge from an investigation by Congress into potentially unethical links between outside companies and researchers at the US National Institutes of Health (NIH) isn't particularly encouraging.

At the request of Congress, the biomedical research agency has been looking into the activities of 81 researchers whose names appeared on lists of consultants provided by biotechnology and pharmaceutical companies, but who hadn't declared their interest to the NIH.

Earlier this month, the NIH's director, Elias Zerhouni, told Joe Barton (Republican, Texas), chair of the House Committee on Energy and Commerce, that about half of the 81 were found to be in breach of the ethics rules that were in force at the time of their consultancy work. Most of the infractions were minor, but eight have been referred to the health department's inspector-general for further investigation.

The steady drip of this sort of information into the public domain since December 2003, when the *Los Angeles Times* first reported a few egregious examples of conflict-of-interest at the NIH, is taking its toll on the public reputations of the agency and its staff.

Zerhouni has moved swiftly to confront the issue. His clampdown on consultancy arrangements and on the holding of investments among thousands of NIH employees has caused much wailing and gnashing of teeth at the agency's main campus in Bethesda, Maryland.

But the rules are being implemented with extended deadlines to allow people sufficient time to alter their financial arrangements.

The clampdown leaves the NIH's intramural staff in a bind, unable to collaborate closely with the biotechnology industry at a time when such interactions have become almost routine for researchers in some sub-disciplines. At some stage, collaboration between researchers and industry must be redeveloped on a basis that will be consistent with the public's reasonable expectations of publicly funded researchers.

Details of the latest batch of infractions haven't been released, but many of them are probably minor, such as meeting an off-site collaborator without requesting a half-day's vacation. Congress is angry because the interactions weren't properly reported under the NIH's previous ethics regime. In some cases, that happened not out of any nefarious intent, but because the NIH is a large and diffuse federation of centres and institutes.

Now Barton's committee wants to centralize the agency. A draft reauthorization bill for the agency would give far more authority to the director's office, and support additional and extensive monitoring and reporting functions there, as well as giving the director more power to enforce cooperation between institutes and centres.

Some reform is due, but this measure goes too far. The NIH needs to modernize, but shouldn't overthrow the autonomy of centres and institutes that has served it so well in the past.

The eventual solution should involve a mixture of self-awareness and common sense. Researchers need to recognize that the conflict-of-interest issue can no longer be brushed off as something for politicians and the press to worry about. The cases that have already been exposed at the NIH amply demonstrate how germane the matter is to biomedical research. ■

RESEARCH HIGHLIGHTS

Here comes the rain

Geophys. Res. Lett. **32**, L13701 (2005)
Climate change may heighten variation in precipitation from year to year, according to Filippo Giorgi and Xunqiang Bi at the Abdus Salam International Centre for Theoretical Physics in Trieste, Italy. Their region-based approach lends weight to previous work that suggests climate change could increase weather variability.

Giorgi and Bi divided the Earth's surface into a grid of regions, each 1 degree square. They used this grid system with 18 different computer models of twenty-first-century climate to see what differences were predicted over the years. In all regions the climate was warmer and the variability in the amount of precipitation from year to year increased significantly.



CANCER GENETICS

Nasty neighbourhood

Nature Genet. doi:10.1038/ng1596 (2005)
Tissue cells surrounding tumours can contain permanent changes to their genes that could encourage tumour development, say researchers. These 'epigenetic' changes take the form of chemical modifications to DNA and are passed on when cells divide.

A team led by Kornelia Polyak at the Dana-Farber Cancer Institute in Boston, Massachusetts, have developed a new way of screening the entire genome of a cell for epigenetic changes. They studied three kinds of cells in the tissues surrounding breast tumours and found alterations in all three. The changes resulted in abnormal gene expression in these cells. The findings suggest that epigenetic changes are involved in creating the abnormal tumour microenvironment thought to foster disease progression.

IMMUNOLOGY

Gut reaction

Cell **122**, 107-118 (2005)
Lack of exposure to harmless bacteria has been blamed for the rising rates of allergic diseases, such as asthma, in industrialized nations. Support for this 'hygiene hypothesis' comes from a study that shows how a sugar produced by a gut bacterium directs the development of immune cells in animals.

A team led by Dennis Kasper of Harvard Medical School in Boston, Massachusetts, demonstrated that mice raised in a germ-free

environment had several immune-system defects. These included unusually high proportions of immune cells called T_H2 cells, whose abnormal activity is linked to allergies. Dosing such mice with the gut bacterium *Bacteroides fragilis* restored normal immune development. The team found this was due to a previously unknown kind of sugar called PSA that is made by the bacterium.

MATERIALS SCIENCE

Cracked it

Phys. Rev. Lett. **95**, 025502 (2005)
When a piece of material breaks, what determines the shape of the resulting pieces? To find out, researchers from the National Centre for Scientific Research in Paris and the University of Manchester, UK, drove a cutting tip through a thin, brittle polymer film.

They say that, under certain conditions, the shape of the crack depends solely on the width

of the cutting tool. It does not depend on the tool's speed, or on the width or thickness of the film. The researchers were also able to reduce cracking behaviour to a simple set of geometrical rules, which they used to reproduce the fracture patterns created by several different shapes of cutting tool.

MEDICAL MICROBIOLOGY

Secret sex life

Curr. Biol. **15**, 1242-1248 (2005)
A fungus that causes life-threatening infections may have been having sex under researchers' noses. Until now, they thought it reproduced only asexually.

Aspergillus fumigatus can cause serious respiratory illness in people with weakened immune systems, and is a major cause of allergies. Paul Dyer from the University of Nottingham, UK, and his colleagues found that the *A. fumigatus* genome contains active genes very similar to those that other fungal species need for sex. They also discovered two different mating types, and evidence that genes can transfer between different populations. If the fungus has been hiding a furtive sex life, geneticists could perform breeding experiments to uncover the genes it uses to cause disease.

SYNTHETIC BIOLOGY

Close couple

Nature Chem. Biol. doi:10.1038/chembio719 (2005)
Scientists have designed a way to make proteins that works independently of the

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

J. BURTON/GETTY IMAGES

normal machinery in a bacterial cell. Jason Chin and Oliver Rackham of the Medical Research Council Laboratory of Molecular Biology in Cambridge, UK, have engineered two key components of this pathway in *Escherichia coli*.

The researchers altered both the 'recipe' for proteins, known as messenger RNA, and the molecular machines called ribosomes that turn this message into a protein. The engineered ribosomes can read only the altered RNA, and this RNA will not work in a normal ribosome. The researchers say these modified ribosome-RNA pairs will allow them to develop sophisticated ways of programming artificial processes inside living cells.

NEUROBIOLOGY

Pore connections

Science doi:10.1126/science.1116270 (2005)
Voltage-dependent ion channels are complex protein structures that open and close, helping electrical signals to travel along nerves. They have sensors that can detect changes of a few hundredths of a volt and make the pore switch between closed and open states. But until now, no one knew for sure how the sensors work.

Roderick MacKinnon and his colleagues from New York's Rockefeller University have determined the first crystal structure of a mammalian channel in its natural state. They found that the voltage-sensing mechanism involves proteins that are quite independent of the pore itself. They also discovered that these sensing regions cross the membrane, which makes them unlike those in other kinds of channels. The team suggests ways in which the sensors could mechanically alter the shape of the pore.

CARBON CHEMISTRY

Get in the ring

J. Am. Chem. Soc. doi:10.1021/ja053202 (2005)

Japanese chemists have made the first stable molecular ring of silicon atoms. Various carbon-ring compounds, such as benzene, contain delocalized electrons that give rings extra stability, but no analogous molecules have been made for carbon's cousin, silicon.

The silicon ring contains three silicon atoms arranged in an equilateral triangle, carrying two delocalized electrons and an overall positive charge. Akira Sekiguchi and his fellow authors from the University of Tsukuba suggest that the rings could be stuck to metals to form catalysts. They now plan to generate all-silicon equivalents of benzene, and even buckminsterfullerene (C_{60}).

ANIMAL BEHAVIOUR

Lighting the way

Biol. Lett. doi:10.1098/rsbl.2005.0334 (2005)

Locust swarms cover great distances, but avoid flying over large bodies of water. The insects (*Schistocerca gregaria*) use the polarized light reflected by the water to steer clear, according to Nadav Shashar and his colleagues at the Hebrew University in Israel.

Light waves reflected from flat surfaces such as water oscillate in a plane parallel to the surface, and some creatures — including locusts — can see this. Shashar's team caught locusts and tethered them above surfaces that reflect light in different ways. Sure enough, the insects tended to fly away from



B. GUZNER

the strongly polarizing surfaces. The researchers hint that such materials could be developed to divert these destructive pests from crops.

MATERIALS SCIENCE

The hard stuff

Chem. Mater. doi:10.1021/cm0505392 (2005)

Polymers that conduct electricity have been toughened up by a team of chemists from the University of Manitoba in Winnipeg, Canada.

The development of polymer-based electronics has been limited by the poor heat-resistance and weakness of the materials. These properties are often due to the chemicals added to 'dope' the material, to boost conductivity. The alternative, linking parallel polymer molecules together to make them stronger, can block the current.

The team has made a polymer based on poly(anilineboronic acid) that is extremely hard and dopes itself. Heating the precursor polymer alters its structure and chemistry, leaving a charged boron atom that both makes the polymer a good conductor and links parallel chains.

JOURNAL CLUB

Adam Summers
University of California, Irvine

A biomechanist bones up on healing processes to work out why sharks are total softies.

Sharks, skates and stingrays have skeletons that are made entirely from a lightly calcified form of cartilage, but their ancestors had perfectly normal bone. A major question in my lab is why this should be. Maybe it's because cartilage is nearly neutrally buoyant, whereas

bone is a real sinker. However, a recent finding has led us to think about the cost of repair too.

Although it is a structural material, cartilage is fundamentally different from bone. It lacks blood vessels and, in mammals, birds and amphibians at least, has virtually no ability to repair itself. Now it seems that cartilage in sharks lacks the ability to heal as well.

Doreen Ashhurst, of St George's Hospital Medical School in London, examined cuts in the fin supports of a shark (*Scyliorhinus*). She found absolutely no healing over six

months (Ashhurst, D. E. *Matrix Biol.* **23**, 15–22; 2004).

I find this very interesting. Ashhurst was testing whether a cartilage repair mechanism had evolved in an animal that relies heavily on the substance. What are the implications that it hasn't?

A human skeleton is shot through with thousands of microfractures in various stages of repair, with osteoclasts dissolving damaged bone and osteoblasts laying down fresh material. This remodelling saves us from developing significant fractures in

bones that go through hundreds of loading cycles every day.

The skeletal elements of a constantly swimming shark may experience more than a billion cycles of loading over a long life. To survive, either the shark's skeleton must be heavily constructed so that it does not deform much on each loading cycle, or its cartilage must be remarkably resistant to fatigue.

My group now wonders whether the cost of continuously repairing the skeleton was a selective pressure that led cartilaginous fish to lose their bones.

NEWS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

METROPOLITAN POLICE/PA/EMPHICS

Deadly plan: the four London bombers arrive at Luton station on their way to the capital's tube network.

Psychologists warn of more suicide attacks in the wake of London bombs

LONDON

When bombers struck London on the morning of 7 July, the city's inhabitants were shocked but not surprised. Since the 11 September attacks in New York, intelligence officials had warned that Britain's capital was a terrorist target. The real surprise came five days later, when police said that they believed the attacks were the work of suicide bombers. The perpetrators were born and brought up in Britain, had normal jobs and — to their neighbours and colleagues at least — had not seemed to be involved with extremist groups.

But to academics who have studied the psychology of suicide bombers and the groups that back them, this was par for the course. Through interviews with bombers who have failed to detonate their devices, and discussions with the families of those who succeeded, reasonable knowledge about suicide terrorism has been accumulated.

From the West Bank to Iraq to Chechnya, studies have shown that — contrary to media portrayals — economic status, educational background and religious beliefs are not significant factors in motivating the bombers. The attackers are often educated and come

from middle-class backgrounds. They are not suicidal in the typical sense, or even depressed.

"It appears that the London bombers fit the profile of suicide bombers more generally in that there is no profile," says Alan Krueger, an economist at Princeton University in New Jersey who studies terrorism. "Suicide bombers come from all walks of life."

Intentional act

Intelligence officers in London are reported to be considering whether the bombers intended to blow themselves up, or could have been duped by terrorist organizers into doing so. If the attackers are confirmed as suicide bombers, Krueger and others say that those investigating the attack — and attempting to prevent a repeat — should ignore the personality of the bombers and focus on the politics of local communities and of terrorist organizations.

Ariel Merari, a psychologist at Tel Aviv University in Israel, has met many thwarted suicide bombers over the past 25 years. His studies show that until now, suicide terrorists have tended to emerge

from societies that back such actions. Many Palestinians, for example, see what they call 'martyrdom' as a legitimate response to the Israeli occupation.

This assent creates volunteers for the second essential ingredient in suicide terrorism: an organization to support the individual attacker. Once involved with these groups, would-be bombers find it difficult to back out without losing face. Some Islamic organizations lock volunteers in by drawing on their faith and the promise of the afterlife, but religion is not always involved. During the Second World War, for example, Japanese military officials used peer pressure and nationalistic pride to persuade pilots to carry out kamikaze missions.

What about the likelihood of further attacks? Robert Pape, an international-relations specialist at the University of Chicago who has studied data on suicide bombings covering two decades, says that the inhabitants of Western cities should brace themselves. He believes that details of the London bombings, including the use of local people who weren't viewed as key

"Logic drives their actions. They're not madmen, they're just playing to a different rule book."



NASA'S DISCOVERY SHUTTLE GROUNDED
Faulty fuel sensor puts launch on hold. Read all about the mission online.
www.nature.com/news

NASA/KSC

S. SENNE/AP

terror suspects, look similar to other attacks linked to al-Qaeda, such as the Bali bombings of 2002. That campaign will continue as the group attempts to oust US troops from the Middle East, says Pape.

But Merari believes there will be no large-scale campaign of the type seen in Iraq and Israel. "The London bombers do not represent the sentiments of their community," he explains. This lack of support makes it harder for terrorist groups to recruit and easier for police to gather information on suspects. As the group leaders usually have to operate from abroad, it also increases the time needed to plan an attack. "This doesn't mean that more attacks will not occur," says Merari. "But there will not be a wave of them."

Calculated risk

Andrew Coburn, director of terrorism research for Risk Management Solutions in Cambridge, UK, uses risk-analysis techniques borrowed from the fields of economics and natural disasters to predict terrorist risk for insurance companies. Understanding that the attackers are, in a sense, sane and rational is key to predicting where they might strike and what damage they will inflict, he says.

"Logic drives their actions," he says. "They're not madmen. They're just playing to a different rule book."

Coburn and his colleagues have developed a terrorism risk model that scores different attacks on the basis of how easy they are to carry out versus how much damage they will inflict — not only in terms of human life, but also in economics and the symbolic value of the target.

"It doesn't predict where attacks will happen, but it suggests the kinds of targets," says Coburn. "There are not many locations in the world where you can guarantee 100 people within four metres of your explosion in one of the least well-defended areas, such as a tube train."

Coburn believes that, after an initial period of particularly high security, when the danger is seen as diminished, the risk to London will remain high for years to come. He does not rule out the possibility of terrorists attempting a far larger attack that, although more difficult to pull off, could claim far more lives than the 56 who perished earlier this month.

This was Western Europe's first suicide attack, and Coburn also sees that trend moving gradually westwards. "The risk is spreading from the Middle East, through Europe and into the United States," he says. But he maintains that London can pull off a safe Olympic Games in 2012. "When we go all out and crank up the protection for an event, people tend not to attack."

Jim Giles and Michael Hopkin

Altered embryos offered as solution to stem-cell rift

WASHINGTON DC

US lawmakers were last week gathering support for legislation that would ease federal restrictions on funding for embryonic stem-cell research.

Public polls favour the bill, which passed by a large bipartisan majority in the House of Representatives on 24 May (see *Nature* 435, 544–545; 2005). But President George W. Bush is dead against it. That left many senators in the Republican party in a quandary: how could they back a popular measure without alienating pro-life voters and the US president?

Into this fray stepped William Hurlbut, a consulting professor in human biology at Stanford University who serves on the President's Council on Bioethics. At a Senate hearing on 12 July, Hurlbut told lawmakers to back research into the creation of embryo-like entities that have been engineered to lack the capacity to develop into human babies, for example by mutating certain genes.

Hurlbut has proposed that scientists create these entities in a process he calls "altered nuclear transfer", to distinguish it from somatic cell nuclear transfer, which scientists use to create human embryos from which stem cells can be extracted.

Such entities lack the moral status of human embryos, argues Hurlbut, and so could be used for research with fewer ethical objections. "We should find a way to go forward with our biomedical research that gathers in our whole nation," he told the Senate.

Although Hurlbut sees himself as a unifying force, many stem-cell researchers are worried. His proposal has not been tested, and the idea of purposely creating defective embryos has met with serious objections from ethicists.

But it is a gift for politicians who are undecided about stem-cell research. Senate leaders, encouraged by the White House, have begun pushing for laws backing such alternatives to embryonic stem-cell research. The proposals threaten to erode support for the measure to loosen funding restrictions on stem-cell research itself.

Despite opposing the use of embryos for stem-cell research, Hurlbut describes himself as "very pro-science". He attended

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

William Hurlbut believes embryo-like entities offer a way forward for stem-cell research.

Stanford Medical School, but abandoned plans to practise medicine when his first child was born with severe brain damage.

As a result, Hurlbut decided to devote his career to teaching and studying the ethics of biomedicine. One of his heroes became Saint Francis of Assisi, whose life of poverty was characterized by a love of nature. "I thought, this is exactly what the world needs right now," Hurlbut says. "We were ravaging the natural world, and it was obvious that we were ramping up to ravage and reorder the human body as well."

Joining the President's Council on Bioethics in 2002, Hurlbut says he felt torn between the supporters and opponents of embryonic stem-cell research, and proposed altered nuclear transfer as a way to bridge the divide.

He warns that the debate over stem cells could be the first of a series of battles over the use of powerful techniques in developmental biology. He believes scientists must forge harmony with their opponents or risk losing support and funding for their work.

"If we don't have a solid frame from which we can go forward, there's just going to be an endless series of bitter disputes," Hurlbut says. "I'm trying to provide one little island of unity in a large sea of controversy."

As *Nature* went to press, negotiators were still trying to decide how to bring the stem-cell legislation before the full Senate for a vote this week.

Erika Check

Bird flu: crossing borders

Despite recent reports from governments that bird flu is under control, it continues to spread through Asia's poultry and claim lives — there are even signs of human-to-human transmission. **Declan Butler** tracks the disease's inexorable spread.

VIETNAM

On 14 July the official Vietnam news agency reported no new human cases of avian flu since 4 June, and said that the government felt infection had been "well-contained". But the same day the *Tien Phong* newspaper reported a further human death, as well as three known and a dozen suspect cases. If confirmed, the death would bring the country's total to 40, with 20 fatalities since the start of the year.

INDIA

In response to the outbreak of H5N1 among migratory birds in China, India announced last week that it will monitor 50 of their arrival points. Bar-headed geese in particular will migrate across the Himalayas in coming months. Blood samples will be taken from birds and tested in a high-security laboratory in Bhopal.

THAILAND

Five new outbreaks in poultry published by the World Organization for Animal Health (OIE) on 15 July signal the failure of a huge government campaign to eradicate the disease, and show that bird flu is now endemic in the country. The outbreaks started on 5 and 6 July, in three districts of Suphanburi province. The country has reported no human cases since last October.

JAPAN

On 11 July, Japan announced the seventh outbreak of H5N2 bird-flu virus in Ibaraki since late June. The H5N2 strain is not yet known to cause illness in humans.

PHILIPPINES

On 8 July, the government announced its first outbreak of avian flu, in ducks living in Bulacan province. Officials were quick to downplay it — agricultural secretary Arthur Yap tucked into chicken at a press conference. On 15 July, the OIE said that tests suggest the virus is an H5 strain that isn't known to cause serious disease. Samples have been sent to Australia for confirmation.

INDONESIA

The president of Indonesia, Susilo Bambang Yudhoyono, asked his government on 17 July to be open about the suspected deaths from avian flu of three members of the same family. "The cause of their deaths must be made clear," he said. "It should not be covered up."

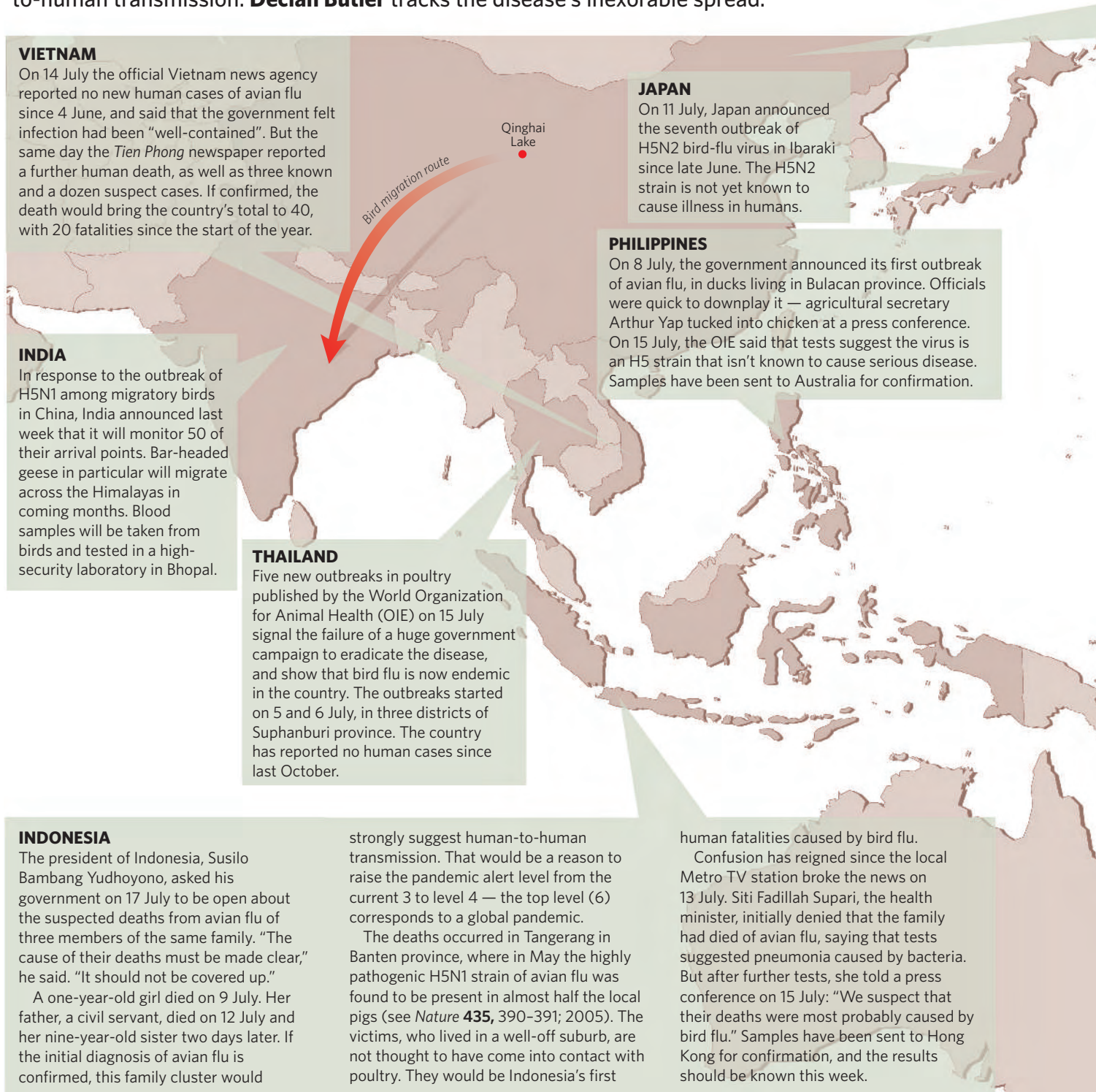
A one-year-old girl died on 9 July. Her father, a civil servant, died on 12 July and her nine-year-old sister two days later. If the initial diagnosis of avian flu is confirmed, this family cluster would

strongly suggest human-to-human transmission. That would be a reason to raise the pandemic alert level from the current 3 to level 4 — the top level (6) corresponds to a global pandemic.

The deaths occurred in Tangerang in Banten province, where in May the highly pathogenic H5N1 strain of avian flu was found to be present in almost half the local pigs (see *Nature* **435**, 390–391; 2005). The victims, who lived in a well-off suburb, are not thought to have come into contact with poultry. They would be Indonesia's first

human fatalities caused by bird flu.

Confusion has reigned since the local Metro TV station broke the news on 13 July. Siti Fadillah Supari, the health minister, initially denied that the family had died of avian flu, saying that tests suggested pneumonia caused by bacteria. But after further tests, she told a press conference on 15 July: "We suspect that their deaths were most probably caused by bird flu." Samples have been sent to Hong Kong for confirmation, and the results should be known this week.



CHINA

Recent headlines from Xinhuanet, China's semi-official news agency, include "Bird flu outbreak in Qinghai 'under control'" and "International organizations 'impressed' by China's commitment in fighting bird flu". They give the impression that China is well in control of the H5N1 outbreaks among thousands of migratory birds at Qinghai Lake in western China, and in Xinjiang province near the border with Kazakhstan (see *Nature* **435**, 542–543; 2005).

But this is difficult to verify because China does not allow free movement of international experts or journalists to the outbreak zones. China's grip on information now looks set to be tightened further through new rules that require all research on avian flu to be vetted by its agriculture ministry.

Concerns came to a head on 8 July, when Xinhuanet quoted Jia Youling, director-general of the agriculture ministry's veterinary bureau, as asserting that a paper on the Qinghai outbreak published online by *Nature* on 6 July "made the wrong conclusion". Jia also accused the authors of never having visited Qinghai, and of carrying out their research illegally because their labs did not meet safety standards and they did not have government approval for the work.

The article's authors included Guan Yi from the Joint Influenza Research Center, run by Shantou University in mainland China and the University of Hong Kong, and Robert Webster, a flu expert at St Jude Children's Research Hospital in Memphis, Tennessee. It concluded that the Qinghai virus is a new, highly virulent form of H5N1, and that the birds risk spreading the virus across Asia as they migrate over the coming months (H. Chen *et al.* *Nature* **436**, 191; 2005). A paper published simultaneously in *Science* reached similar conclusions (J. Liu *et al.* *Science* doi:10.1126/science.1115273; 2005).

Jia objected in particular to an additional conclusion in the *Nature* paper that the isolates were similar to ones "isolated from poultry markets in Fujian, Guangdong, Hunan and Yunnan provinces during 2005". China has not declared any avian flu outbreaks in poultry to international authorities this year. "The article's conclusion lacks credibility," Jia is reported as saying. "No bird flu has broken out in southern China since the beginning of this year."

Guan refutes Jia's allegations, saying that his lab meets the World Health Organization's standards for biosafety and collaborates with flu experts around the world. He reiterates that his team found H5N1 in samples taken from poultry in the region this year.

Guan interprets Jia's stance as one more example of government 'pressure' on scientists trying to investigate the country's flu outbreaks. On 16 June, the Chinese agriculture ministry warned that it would "regulate and investigate research and testing without permission, to stop unauthorized work". This warning follows a series of rules it published on 31 May, requiring scientists to apply for permission to collect and study H5N1 samples, and to have their results double-checked by the ministry.



CARBON DATING WORKS FOR CELLS

Radioactive fallout from nuclear tests serves as measuring stick.

www.nature.com/news

Arsenic-free water still a pipedream

Decontamination plants installed at wells throughout West Bengal are failing to reduce arsenic in local drinking water to safe levels, according to a report. The authors suggest that efforts to supply safe water should instead focus on purifying surface water.

Of 18 arsenic-removal plants monitored over a two-year period, none reduced arsenic levels below the maximum safe value stipulated by the World Health Organization (WHO), says epidemiologist Dipankar Chakraborti of Jadavpur University in Calcutta, India, whose team carried out the tests (M. A. Hossain *et al. Environ. Sci. Technol.* **39**, 4300–4306; 2005). The findings come as a blow to efforts to address what has been called the worst mass poisoning in history, in which millions of people were exposed to dangerous or fatal levels of arsenic in their water.

The high levels of arsenic in well water used for drinking and irrigation came to light in the early 1990s, after outbreaks of skin disease and cancers in West Bengal and Bangladesh. The arsenic comes from natural geological sources that weren't recognized when the wells were dug during the 1970s. An estimated 35 million people were drinking from such wells, dug by aid agencies so that locals wouldn't have to rely on rain and river water, which is often contaminated by carriers of diseases such as typhoid and dysentery.

To try to fix the situation, some 2,000



Hard to swallow: millions of people in West Bengal are still drinking highly contaminated water.

arsenic-removal plants were installed in wells in West Bengal, and many more in Bangladesh, at an average cost of US\$1,500 each. These plants aim to remove arsenic from water using a series of filters and extraction systems.

But controversy has persisted. Chakraborti

has previously shown that methods for identifying safe wells are not always reliable. Now he says that the removal plants cannot be trusted either.

Chakraborti and his colleagues tested 18 such plants, from 11 different manufacturers in India, Germany and the United States. The average arsenic concentration in water treated during a two-year period was 26 micrograms per litre — more than twice the value recommended by the WHO. Only two of the plants met the Indian standard value for arsenic levels, which is five times higher than that of the WHO, and 80% of the local villagers tested had abnormal levels of arsenic in their urine.

Chakraborti believes that efforts should now focus on harnessing and purifying surface water. The United Nations Children's Fund (UNICEF), which initiated and funded the original well-drilling programme, supports the approach. "We're really pushing for rain-water harvesting," says UNICEF emergencies coordinator Paula Plaza.

However, there are now plans to test arsenic-removal plants more thoroughly before they are licensed for use. A Canadian company called the Ontario Center for Environmental Technology Advancement is collaborating with the Bangladeshi government to develop performance standards for these systems. ■

Philip Ball

I. BERRY/MAGNUM PHOTOS

Malaysia plans 'red book' in its attempts to go green

KUALA LUMPUR

Malaysia, criticized in the past for being a poor steward of its biodiversity, seems to be turning over a new leaf.

Biodiversity experts from across the country met in Kuala Lumpur last month to hammer out a plan that would catalogue the country's thousands of plant and animal species.

Malaysia is one of the most biologically diverse countries in the world. It is thought to host around 15,000 different plant species, although only about half that number have been found and listed. But the country has become infamous in recent decades for clearing rainforests and draining peat swamps to grow palm trees — it now produces around half of the world's supply of palm oil. The effect of this on the country's biodiversity is not known.

The international Convention on Biological Diversity has been pressuring Malaysia to come up with conservation strategies since it came into force in 1993, with little result. But Dato' Seri Abdullah



Malaysia's drive to grow palm trees for palm oil has had untold effects on the country's biodiversity.

Ahmad Badawi, who became prime minister in 2003, is said to be more sympathetic than his predecessor to ecological issues. In particular, he is keen to use the country's biodiversity to drive drug development.

That momentum, and associated funding, has led to a project to create a national 'red book'. This would catalogue which species are present and where, as well as listing any threats they face. The project will be overseen by Saw Leng Guan of the Forest Research Institute of Malaysia near Kuala Lumpur.

But Peter Ng, a conservation biologist at the National University of Singapore, says Malaysia will be a tough nut to crack because knowledge is so fragmented. Drastic funding cuts in the 1990s also left the country with few taxonomists. "We have money, but we need people to do the work," he says. ■

David Cyranoski

M. EDWARDS/STILL PICTURES

Asia squeezes Europe's lead in science

WASHINGTON DC

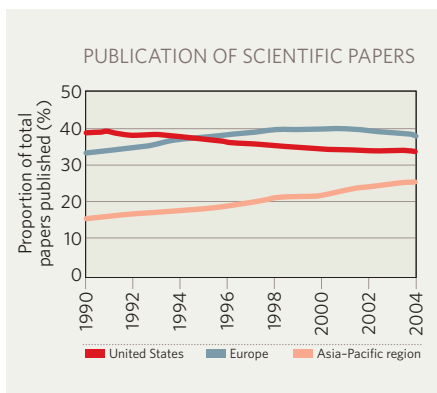
Asian nations are catching up with Europe and the United States in terms of scientific output, says a US report. If current trends continue, publications from the Asia-Pacific region may outstrip those from the United States within six or seven years.

In 2004, the report shows, countries from the Asia-Pacific region, including China, South Korea, Taiwan, Japan, Singapore and India, produced 25% of the world's research papers. In 1990, Asia's share of the scientific output was just 16%.

"The rise of the Asia-Pacific nations seems the most striking thing," says Christopher King, editor of the newsletter *Science Watch*, which published the report in its July/August issue. The newsletter is produced by Thomson Scientific, based in Philadelphia, Pennsylvania, which owns a widely used citation database — Thomson ISI.

By comparison, Europe last year produced 38% of the world's scientific papers, and the United States produced 33% (see Graph). Although it is the current world leader, Europe is beginning to worry. The European Commission is due to release a report this week saying that the European Union (EU) may not reach its stated spending goals for research and development by the end of this decade.

Thomson's findings echo a highly regarded



2004 National Science Foundation (NSF) analysis — the biennial Science and Engineering Indicators. This showed that the number of US papers published has remained essentially flat over the past decade, whereas the rest of the world has been publishing more with every year.

Within Asia, the NSF analysis found, China, South Korea, Singapore and Taiwan grew the most. Between 1988 and 2001, article output rose nearly fivefold in China, sixfold in Singapore and Taiwan, and by 14 times in South Korea. At the same time, article output rose only 1.1 times in the United States, 1.6 times in Europe and 1.4 times worldwide.

Those trends haven't changed much in the

past three years, says Robert Bell, a senior analyst at the NSF.

One reason for the higher Asian publication share is strong economic growth and the resulting increase in research funding, says Mu-Ming Poo, a neuroscientist at the University of California, Berkeley, who spends part of every year as director of the Institute of Neuroscience in Shanghai.

What's more, Poo says, research performance in Asia is now increasingly evaluated in terms of the publications in journals that are indexed by Thomson Scientific.

In China, some institutions even pay researchers extra for publications in indexed journals, especially ones that carry widely cited articles, says Wu Yishan, who analyses Chinese research performance for the Institute of Scientific and Technical Information in Beijing. "Such incentives are effective in promoting more publications," he says.

The reasons behind the stagnating number of US publications are less clear, says Bell. The NSF has collected data from the top 200 US universities to look for correlations between the number of research publications and other factors, such as the number of postdocs and graduate students or the amount of research funding. All these factors are important, says Bell, but no particular one stands out as the driving force behind scientific output. "There is no smoking gun," he says.

As for Europe, its first-place standing does not mean everything, cautions Vincent Duchêne, an analyst at the European Commission in Brussels. The United States publishes more papers per researcher than Europe, and with greater impact.

Duchêne helped to prepare this week's EU report, which warns that Europe may fall further behind. In 2002, the EU set a target for its research and development 'intensity': 3% of its gross domestic product by 2010. In 2003, its intensity was 1.9% — lower than the United States' 2.6% and higher than China's 1.3%. But China's intensity has grown by more than 10% a year, while Europe's has increased only gradually.

If those trends continue, the report says, China's research and development intensity will catch up with Europe's by 2010.

Andreas von Bubnoff

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Paper tigers: countries in the Asia-Pacific region are boosting their production of scientific articles.

L. FRITZ/NOAA FISHERIES SERVICE/PERMITS: 782-1532-00; 782-1532-01; 782-1532-02



A programme to save Steller sea lions has been criticized for using hot brands on pups.

Animal-rights group sues over 'disturbing' work on sea lions

Fresh allegations have surfaced about an extensive US research programme designed to save endangered sea lions. The charges say it may actually be harming the species.

On 12 July the Humane Society of the United States sued the National Marine Fisheries Service (NMFS), saying that it failed to properly conduct and monitor studies on Steller sea lions (*Eumetopias jubatus*) in Alaska.

The lawsuit alleges that the NMFS violated animal-protection laws "by issuing multiple research permits to a wide variety of entities that allow intrusive, duplicative, uncoordinated and unnecessary research on Steller sea lions".

The records of the US Marine Mammal Commission, an independent panel that oversees research, partly support the society's allegations. William Hogarth, a top official at the NMFS, wrote in a letter to the society on 22 June that he also has concerns.

Hogarth has ordered an extensive environmental study on the Steller research programme, which is expected to take two years and may cost up to \$500,000. The NMFS has pumped more than \$120 million over the past four years into research on why Steller populations have plunged in the past quarter-century (see *Nature* 436, 14–16; 2005).

Administrators of the Steller research programme say the studies are conducted appropriately, with only one or two animal deaths annually. "No one is doing anything that is not important for research," says Douglas DeMaster,

an NMFS biologist and director of the Alaska Fisheries Science Center in Seattle.

In May and June the NMFS issued research permits before a review period required by the Marine Mammal Commission was complete. "This is particularly disturbing in light of the scope of the proposed research and potential for adverse effects," wrote David Cottingham, the commission's executive director, in a letter to the NMFS on 10 June.

The commission has regularly questioned aspects of Steller research, citing the 'hot branding' of hundreds of pups, the sedation of animals, and the lack of follow-up reports on previous studies, among other factors.

One researcher seeking a permit this year was Randall Davis, a physiologist at Texas A&M University, Galveston. On 31 May, Davis was cited for alleged research violations in 2003 and 2004. He faces a \$10,000 fine and research restrictions after his team reportedly used an unapproved sedation drug and captured Stellers in violation of his research permit.

Davis is contesting the charges, and an administrative hearing is planned. He told *Nature* that some NMFS officials were "completely out of control", adding that they were "hostile to researchers".

NMFS officials say they are discussing the concerns with the Humane Society. A society attorney says that if no remedy turns up, he will seek a court order to suspend some projects. ■

Rex Dalton

ON THE RECORD

"All I can say is 'shucks'. We ran out of gas."

Wayne Hale, deputy manager of the US space shuttle programme, reacts to the postponement of the shuttle's launch due to a faulty fuel sensor.

SCORECARD



Big spender

Illinois governor Rod Blagojevich joins the biotech bandwagon with a decision to spend \$10 million on stem-cell research.



Fight club

Thai officials plan 'passports' for fighting cocks in effort to track bird flu's spread through pugilistic poultry.



Where's the beef?

Reality bites for Canadian cows as a court ruling overturns objections to US plan to import beef from up north.

NUMBER CRUNCH

898 miles How far the average UK resident drives every year to shop for food.

19 million tonnes The amount of carbon dioxide emitted in Britain in 2002 transporting food to the dinner table.

\$15.5 billion The minimum annual cost to Britain of these 'food miles'.

Source: UK Department for Environment, Food and Rural Affairs.

OVERHYPED

Alcoholism treatments

'Talk' therapies, such as the 12-step programme, are commonly used to help alcoholics give up their vice — but do they really work? A 1997 study suggested that they offered unequivocal benefits, but a reanalysis of the original data is casting a somewhat different light.

Although alcoholics who finished a course of treatment did well, those who dropped out of the course before it started were almost as successful at curing their cravings. Those who did worst were those who quit therapy after a single session, suggesting that secret to success is more will power than therapy.

Weapons task force sets sights on single US lab

The United States should consolidate its nuclear-weapons production facilities at a single centralized location, according to a Department of Energy panel. But all the nation's nuclear-weapons labs appear safe for the time being, as department officials are only now beginning to study the recommendations.

A six-member panel led by physicist David Overskei, president of a San Diego company called Decision Factors, released its report last week in draft form. It calls for all nuclear-weapons materials to be located at one site, where production, manufacturing, assembly and disassembly of weapons would all take place. There is no suggestion in the report of where that single facility should be located. Those jobs are currently scattered among several laboratories.

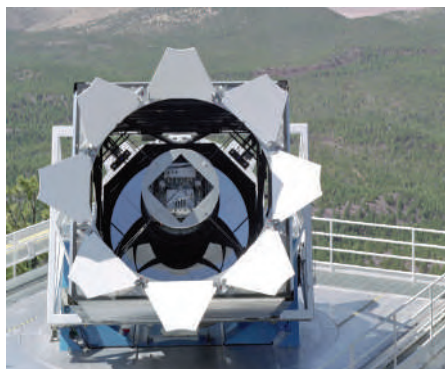
Any plan to move tasks away from existing labs would face an uphill battle in Washington, where congressional representatives from states such as California and New Mexico already fight each other for scarce energy funding.

Sky survey looks to starry future after cash boost

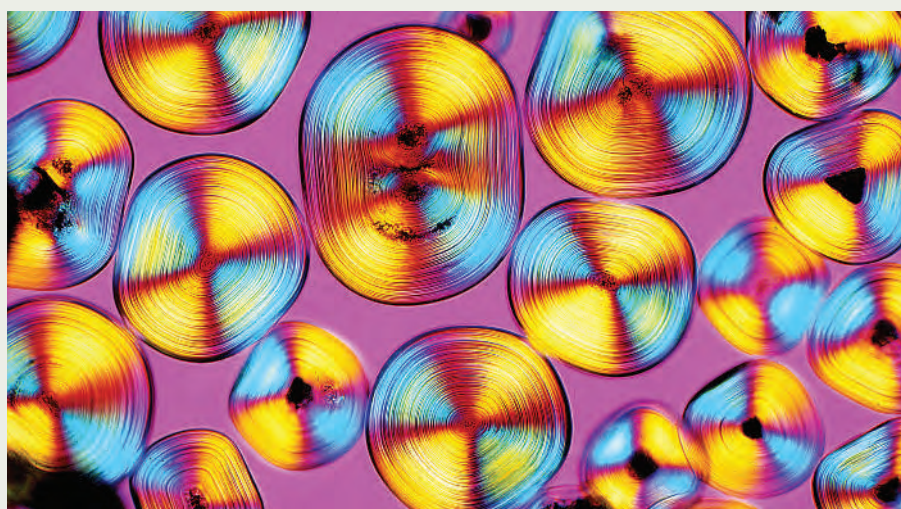
An astronomical project to map millions of celestial objects has received fresh funding to extend its mission for another three years.

The Sloan Digital Sky Survey (SDSS) surveys the night sky using a telescope at Apache Point Observatory in New Mexico. It has already given astronomers new insights into the three-dimensional structure and early make-up of the Universe.

A consortium of funding agencies in the United States, Japan and Germany has now awarded the project \$14.9 million over three years. The money will be used for the second phase of the project, SDSS-II, to study the structure and origins of the Milky Way, and the nature of dark energy.



Looking up: the Sloan Digital Sky Survey will now focus on the structure of the Milky Way.



Kidney stones add colour to scientist's image

A psychedelic beauty belies the pain that can be caused by these objects — common kidney stones, as seen through a microscope.

The concentric rings are layers of calcium oxalate that crystallized one after another, causing the stones to grow. The image is one

of several being exhibited at the 2005 Biomedical Image Awards at the Wellcome Trust in London.

The stones shown here are 0.2 to 0.35 micrometres across, and are thought to come from an animal, not a human.

Japanese law tackles welfare of lab animals

Japan's parliament has updated a law that aims to strengthen the protection of laboratory animals.

The current animal-welfare law, which dates from 1973, requires that researchers try to alleviate the suffering of animals — but it doesn't mention reducing the number of animal experiments or using alternatives where possible. The updated version includes all three key ethical principles, which are widely endorsed by the international community.

Many see the revisions, which are expected to come into effect by June 2006, as a step forward in clarifying Japan's position on animal research. But animal-rights activists point out that researchers will not be required to include the two new principles, only to consider including them.

The Japanese government also plans to set guidelines on how to conduct animal tests. Researchers are creating a third-party accreditation system of labs by monitoring each other's work or asking a US organization to oversee the tests.

China gears up to relaunch human space flight

Since Yang Liwei's historic trip to space in 2003, no Chinese astronaut has orbited the Earth. But that could be about to change, with the scheduled launch of a second mission this autumn.

The Shenzhou VI spacecraft will carry two astronauts into space in early October for five or six days, government officials told the *China Daily* newspaper last week. The astronauts will be chosen from a group of fighter pilots.

China has also spoken of launching at least two meteorological satellites before the 2008 Olympics in Beijing, to provide better weather forecasts for the event.

Europe names advisory board for research council

Some months before the proposed European Research Council (ERC) is given the expected official go-ahead, the European Commission has released the names of 22 high-level scientists, from 17 different countries, who will help to shape it.

Appointees to the newly formed Scientific Council include gene therapist Claudio Bordignon of the San Raffaele Institute in Milan, Italy; atmospheric chemist Paul Crutzen of the Netherlands, joint winner of the 1995 Nobel Prize in chemistry; Fotis Kafatos, former director of the European Molecular Biology Laboratory in Heidelberg, Germany; astrophysicist Maria Teresa Lago of the University of Porto, Portugal; and zoologist Robert May, president of the Royal Society in London.

The group is meant to serve as an independent advisory board to the ERC. If all goes as planned, the research council will serve as the first European-wide granting agency for basic research, beginning operations in 2007.

Just around the corner

For 50 years, physicists have been promising that power from nuclear fusion is imminent. Now they are poised to build an experiment that could vindicate their views.

But will the machine work?

Geoff Brumfiel investigates.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

When word came last month that a site had finally been chosen for the international fusion experiment ITER, Gerald Navratil summed up his feelings in a single word: “relief”. Navratil, a plasma physicist at Columbia University in New York and a member of the US ITER team, had spent the past 18 months on the sidelines, watching helplessly as France and Japan fought over which of them would host the machine.

As the battle raged on, he had seen US support for the project slowly slip away. But with the location for the reactor now fixed in Cadarache, southern France, Navratil is feeling more positive. “Now that we have a site, we can finally proceed in taking the next step,” he says.

Fusion research is difficult enough without the politics. The much-vaunted concept of mimicking the Sun and generating power from nuclear fusion has been an unfulfilled promise for some 50 years. Even to get this far, researchers have had to overcome formidable technical and scientific barriers. But they hope that ITER will at last prove to doubting politicians and scientific colleagues that nuclear fusion is a viable energy source. If all goes well, funding from ITER's six international partners — China, the European Union, Japan, South Korea, Russia and the United States — could be in place this winter, allowing construction to begin in 2006, and operation in 2016.

ITER is designed to heat hydrogen to hundreds of millions of degrees centigrade, and then squeeze energy from the resulting plasma, while holding it stable for minutes at a time. Although most fusion researchers agree

that the reactor will probably be able to generate more power than it consumes, there are some who believe it may struggle to produce as much energy as predicted, or to hold the plasma stable for as long as hoped. “Like any good scientific experiment, there's a chance that it won't work,” says William Dorland, a theoretical plasma researcher at the University of Maryland in College Park.

But ITER is not just any scientific experiment. With construction costs of US\$5.5 billion, it will be one of the most expensive scientific facilities ever built on Earth. “ITER will do what it's supposed to do and give renewed credibility to the field,” asserts Richard Hazeltine, a plasma physicist at the University of Texas at Austin and head of the Fusion Energy Sciences Advisory Committee for the US Department of Energy.

Collision course

Fusion is a simple idea that is hard to achieve in practice. Unlike fission, which generates power from the decay of heavy atomic nuclei, fusion occurs when lightweight nuclei, usually from hydrogen, collide with each other and fuse together to form a new element, typically helium. Those collisions are difficult to orchestrate because the positively charged

nuclei must overcome their natural repulsion. This only occurs when they are moving very fast or are packed closely together. In other words, you need a very hot, dense plasma to achieve fusion power.

The nearest natural source of fusion energy is the Sun. Within its core, gravity pulls positively charged hydrogen nuclei together until they become hot enough and dense enough to fuse into helium. But here on Earth, more inventive solutions are needed.

Shortly after the Second World War, scientists in the United States and Russia began to work on machines that might be able to heat and pressurize hydrogen by squeezing the atoms within strong magnetic fields. Scientists experimented with many different machines, each with their own advantages and quirks, but by the 1960s their search had narrowed towards a single design: the tokamak.

Developed by Russian physicists Andrei Sakharov and Igor Tamm, the tokamak is a doughnut-shaped machine that uses a series of overlapping magnetic fields to hold a hot, dense plasma within the reactor walls (see graphic). Throughout the 1970s, tokamak technology advanced rapidly, leading many to believe it could be fashioned into a prototype power plant by the start of this century.

Those predictions proved to be too optimistic. The energetic hydrogen gas that tokamaks were meant to contain defied confinement by leaking out of its magnetic shackles. Researchers found themselves battling poorly understood plasma instabilities and turbulence, and made slow progress. Even today only one machine, the JT-60 tokamak in

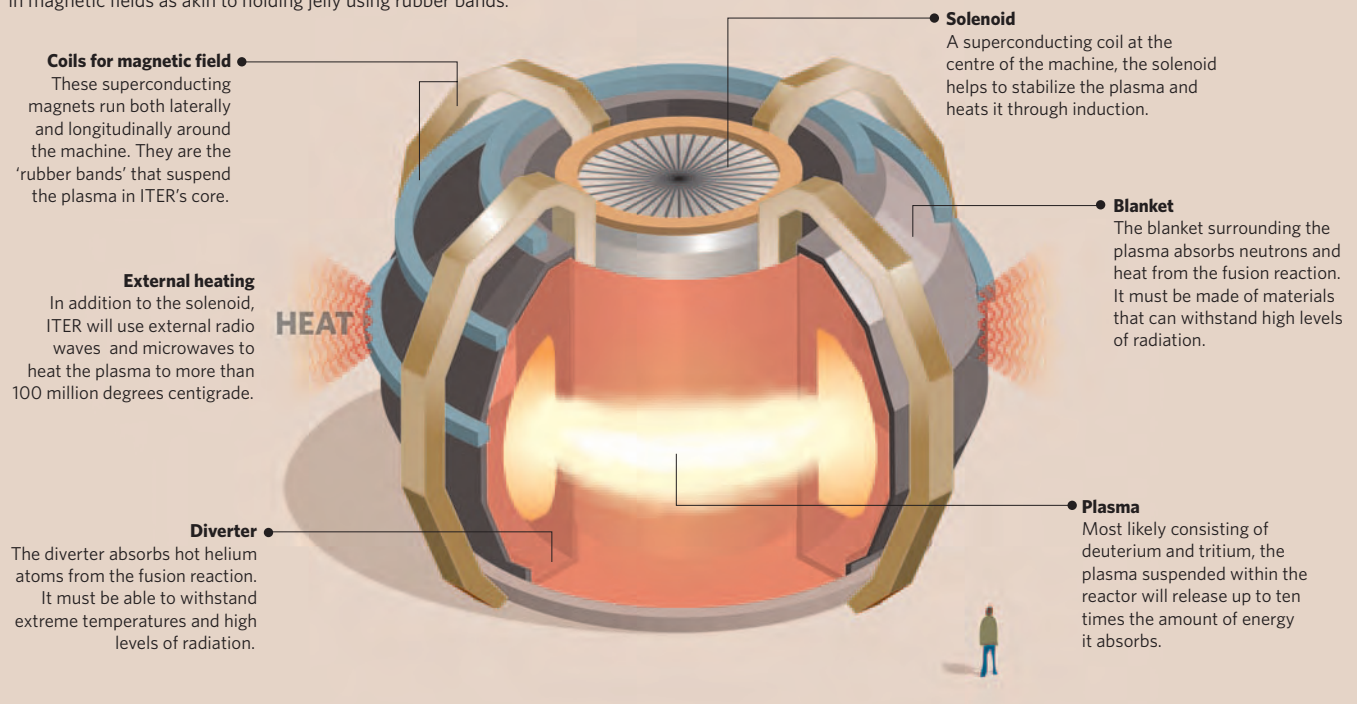
“Fusion science is on the edge of vanishing — we need to go ahead and turn this damn thing on.”

— Richard Hazeltine

B. HORVAT/AFP/GETTY IMAGES

ITER'S TOKAMAK — TOO HOT TO HANDLE

Fusion scientists often describe the job of containing a hot plasma in magnetic fields as akin to holding jelly using rubber bands.



Naka, Japan, has begun to approach the 'break-even point' at which as much energy comes out of the device as goes into it.

Nevertheless, when ITER was first proposed in 1985, the tokamak design was the obvious choice, says Roberto Andreani, who directs the technical efforts of the European Fusion Development Agreement in Garching, Germany. "In all the years that we have studied fusion, I would say that the tokamak has been the most reliable," he says. "It is the only reasonable choice for ITER."

Fuel for thought

ITER's ambitious goal is to hold its hydrogen fuel (a mixture of deuterium and tritium, two isotopes of hydrogen) tightly for between seven and fifteen minutes, while heating it to more than 100 million degrees centigrade and squeezing out about 500 megawatts of energy. The current world record for a sustained high-temperature, high-pressure plasma, held by the JT-60, is 24 seconds. To reach its goals, ITER will use superconducting magnets 25% stronger than those in the Japanese machine, and a host of external heating techniques.

ITER's other main advantage will be its size, says Raymond Fonck of the University of Wisconsin at Madison, a member of the ITER design team. Put simply, the more space a hot plasma has to roam, the better it will behave, Fonck explains. With an outer radius of 6.2 metres and a plasma volume of 840 m³, ITER will be twice as big as any previous tokamak.

But even with its high magnetic fields and enormous size, the machine faces some serious challenges, says Dorland. Unlike previous

machines, ITER is designed to release more energy than it takes in, he says. In the past, researchers have been able to control the plasma temperature simply by turning down the heat. But if ITER succeeds, the plasma will burn under the power of its own fusion reactions, and that means researchers will have to learn how to manage the power output. "The name of the game is to hold a lot of energy into a small place and let it out in a controlled fashion, and that's not easy," he says.

Most of the energy released by ITER will be

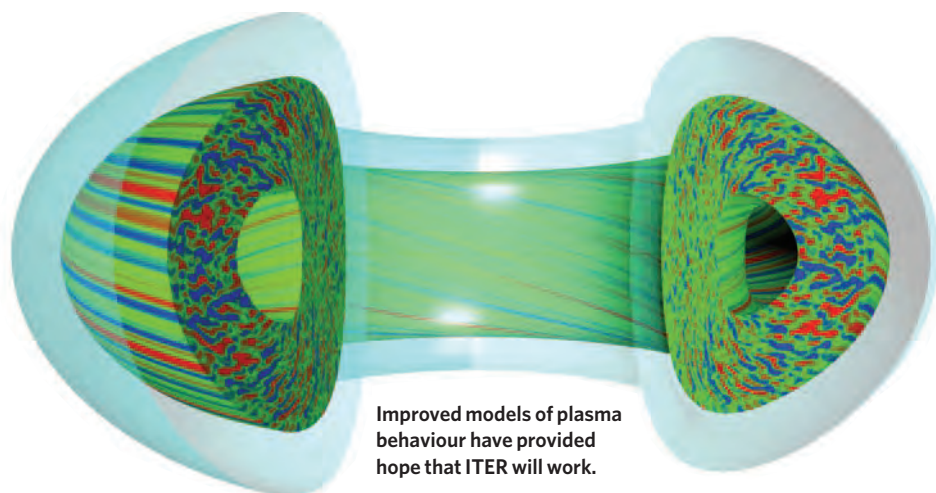
in the form of fast-moving neutrons, which will irradiate the beryllium-coated blanket surrounding the plasma. Hot helium atoms, the other by-product of fusion reactions, will be captured mainly by the diverter, a carbon-coated structure sitting at the bottom of the tokamak. Not everyone agrees that beryllium — used in military armour — and carbon are suited to handling the intense heat and radiation that ITER will release, says Navratil. Tungsten is the likely candidate to replace both in a demonstration power plant to be built after ITER. To aid the search for better structural materials, the Japanese may build a testing facility (see 'Material gains in the east', overleaf).

At some point, a commercial fusion reactor will need to find a way to use the excess heat that is generated. Although ITER is not designed to do this, future machines will be expected to produce useful energy. Most likely, liquid coolants will be used to cool the blanket and diverter, and the hot coolant leaving the reactor will be used to heat water, which in turn will drive steam turbines.

A big uncertainty with ITER's design is the proliferation of helium inside the reactor. Previous tokamak experiments have produced small amounts of fusion, and so small amounts of helium and neutrons. But this time, ITER will fill with substantially more hot helium, and how that will interact with the rest of the plasma could be unpredictable, says Stewart Prager of the University of Wisconsin and another US physicist on the ITER design team.

Making plans: former French prime minister Jean-Pierre Raffarin gets a tour of ITER at Cadarache.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS



The researchers also worry about the walls accumulating too much radioactive tritium, on which there are strict limits for safety reasons.

It is questions like these that caused ITER to scale back its original plan. The machine was initially slated to be twice as big again, capable of holding burning hydrogen for several hours at a time and releasing some 1,500 megawatts of power. But governments were uneasy about the projected costs, and that, together with technical uncertainties raised by Dorland and others, caused a redesign with more modest goals and a smaller price tag.

Some of the technical concerns resulted from numerical simulations of plasma turbu-

lence, which have greatly improved over the past ten years. But uncertainty remains about how hot the outside of the plasma will be, says Dorland, and this will ultimately affect how well ITER functions.

Heated debate

Partly because of better modelling, most researchers agree that the smaller, cheaper ITER can be made to work. "In Europe we're optimistic about the future of fusion," says David Ward, a plasma physicist who studies the economics of fusion for the United Kingdom Atomic Energy Authority in Culham near Oxford. "But the big step will be going

from where we are to ITER." Ward says that the reactor will provide definitive proof of the value of fusion power, as well as offering technical information about how to build a first-generation commercial reactor.

Dorland agrees, with one caveat: "I think that ITER will work, but I'm willing to bet you \$100 that another fusion device will get more power out before it does," he says. There are perhaps half-a-dozen designs that might catch up with tokamaks, but the most impressive to date, he says, is a variation of the tokamak called the spherical torus. This more closely resembles a pitted apple than a doughnut, a shape that allows it to create a sharp boundary between the hot hydrogen plasma and the outer wall of the reactor and squeeze the plasma more tightly. Such a design might achieve a burning fusion reaction with less fuss than the more cumbersome tokamak, he says.

Even so, those who have been in the fusion business a long time believe that it is better to go ahead with ITER than to hope another device will get there first, says Hazeltine. Decades of promises and billions in investment have left international fusion research in what he describes as a fragile condition. "Fusion science is on the edge of vanishing," he says. "I think we need to go ahead and turn this damn thing on."

Geoff Brumfiel is *Nature's* Washington physical sciences correspondent.

J. CANDY/GENERAL ATOMICS

Material gains in the east

With Europe ploughing all of its fusion funds into the ITER experiment in France, the Japanese fusion community is poised to take the lead in several supporting projects. Japan could build facilities related to materials testing, upgrade its JT-60 tokamak (pictured) for new plasma experiments, and host a design centre for DEMO, the demonstration power plant that will follow in ITER's wake.

Notable among these candidate projects is the International Fusion Materials Irradiation Facility (IFMIF). This would aim to verify whether key structural materials, including the blanket (see graphic), can withstand the onslaught of high-energy neutrons that a fusion reactor's plasma will throw at them. IFMIF will require a linear particle accelerator to generate neutrons with energies of about 14 mega-electronvolts and at intensities orders of magnitude higher than existing sources.

The ability of materials to withstand the high-energy neutrons created by ITER and DEMO has never been tested at

such high intensities, and the IFMIF experiments are essential to prove to safety regulators that these machines will survive such extreme conditions, says Yoshikazu Okumura, a fusion scientist at the Japan Atomic Energy Research Institute (JAERI) in Naka. "At the moment, we just don't have any data," he says.

High-energy neutrons can displace atoms from their positions in the structural materials and eventually weaken them. ITER is designed to experience short-term pulses of the fusion reactions, so structural damage should be minor. In DEMO, which would need to operate continuously for years, the accumulated toll of these destabilizing neutrons is likely to be more than 30 times what it will be for ITER.

The current choice for ITER's main structure is stainless steel, whereas DEMO may use a heat-resistant 'ferritic steel', which will endure radiation more readily. But as fusion scientists aim for higher efficiency — and so higher temperatures — new materials, such as silicon carbide, might be



JAERI

needed. "Every country has its own view on how to balance efficiency with safety," says Hideyuki Takatsu, deputy director of JAERI's fusion-technology division.

But at ¥310 billion (US\$2.6 billion) for the 40-year project — roughly one quarter of ITER's price tag — it is not certain whether IFMIF will ever be built, nor whether it would be in Japan, says Takatsu. Other ITER partners, such as Britain, are keen to host IFMIF, and with some expressing frustration at the financial deal

secured by the Japanese for its future role in ITER (see *Nature* **435**, 1142–1143; 2005), the negotiations over IFMIF could become fractious.

But the facility is needed to give a boost to materials science, a field that many fusion researchers say has been neglected. "Materials research has had only about 1–2% of the fusion budget," says Hideki Matsui, a materials scientist at Tohoku University. "There's been an imbalance."

David Cyranoski

¿Vive la revolución?

Cuba's socialist science policies are producing top-notch research from scant economic resources. But, as **Jim Giles** reports, they have harsh consequences for scientists who do not fit in with government priorities.

It's mid-afternoon in Havana's cavernous convention centre, and Cuba's leading scientists are extolling the virtues of the revolution. Cuba's vaccine programme, says one speaker, is the fruit of socialism. Another tells us that the revolutionary leaders have saved the country's environment. Behind him, and not for the first time this afternoon, the giant screen is filled with an image of the commander-in-chief, the bearded one: Fidel Castro.

It's classic propaganda, of course. But this impoverished Caribbean nation does punch above its weight in science, boasting achievements such as the world's only effective vaccine against meningitis B. Despite suffering decades of crippling US sanctions (see 'Neighbourhood dispute', overleaf), and an economic meltdown since the collapse of the Soviet Union, Cuba has achieved first-world levels of education and trained a skilled scientific workforce. But it has done so while restricting the ability of scientists to work in other countries — a freedom that academics in many other nations take for granted.

For an outsider, it is a strange and confusing environment. "You'll never understand Cuba," jokes William Edmundson, director of the British Council in Havana, who has organized numerous UK–Cuban science exchange programmes. "I'm much more relaxed now that I've given up trying."

But after spending a week visiting the country's research institutes, the logic that underpins Cuban science begins to fall into place. Government-funded science is more like a corporate research programme than an academic pursuit; scientists' individual interests are subservient to goals determined from above. But instead of being driven by profits, these goals are set according to the social priorities of Castro's revolutionary government. With Cuba's economy in desperate trouble, this approach has increasingly concentrated resources on applied biomedical, environmental and agricultural projects,

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

leaving basic research out in the cold.

Cuba's science model has its roots in the 1959 revolution. Over the 30 years that followed, Castro built strong links with the Soviet bloc, sending young researchers to the Soviet Union for training. But unlike Soviet science, which had a strong bias towards projects that strengthened the military-industrial complex, Cuba focused on health and social benefits. "We combined applied and basic research, but all of it was for the good of society," says Pedro Valdés Sosa of the Cuban Neuroscience Center in Havana, who helped to establish the country's first brain research lab in 1970.

Dire straits

When the Soviet Union collapsed in 1991, Cuban science reached a crunch point. Cuba suddenly lost its biggest trading partner and source of economic aid. During the 'special period' — the government's euphemistic term for the country's crisis during the early nineties — about a third was sliced off the nation's gross domestic product (GDP).

Rather than letting all Cuba's labs suffer equally, Castro's government chose to continue investing in applied research projects, while effectively allowing basic research to wither on the vine. Money continued to flow into the Western Havana Scientific Pole — a leafy suburb that is home to 50 or so mostly applied research centres and their industrial offshoots. Official statistics are hard to come by, but around US\$1 billion is thought to have been invested in the Scientific Pole during the 1990s. It was to celebrate the fortieth birthday of one of the pole's institutes, the National Center for Scientific Research, that the convention-centre conference was organized.

The Scientific Pole can claim some impressive biomedical achievements: in addition to the meningitis B vaccine, it has produced a cancer vaccine that, despite considerable opposition from anti-Castro politicians, has been licensed for use in the United States. And about two dozen foreign drugs firms are considering exploiting other Cuban products, says George Morris, chief operating director of the London-based Beckpharma, which commercializes drugs developed in academic institutes. Not bad for a nation whose GDP per capita is around a tenth of the European average, and where scientists earn just a few hundred dollars a month.

Standing in a muddy field in wellington boots and a grubby sleeveless top, Osvaldo Franchi-Alfaro Roque embodies another Cuban enterprise that is attracting foreign interest. Although agricultural research hasn't enjoyed the same high-tech success as biomedical research, it is noted for its innovation and environmental friendliness — both enforced through economic necessity. Franchi collaborates extensively with the University of Agriculture of Havana, and his farm in San José is like an open-air inventor's workshop: his device to control irrigation flows, built mainly from a



J. GILES/NATURE

Green glade: ecotourists and a forestry project provide income for people living in Las Terrazas reserve.

plastic bottle and parts scavenged from old cars, has been adopted by farmers across Cuba and neighbouring countries. A large tank next to his avocado trees contains homemade organic pesticide — a mixture of water and local natural products, including seeds from the neem tree (*Azadirachta indica*).

This make-do-and-mend approach is a Cuban tradition, but became vital during the special period. Franchi and other small farmers were forced to experiment with organic pesticides and fertilizers, thanks to the collapse of agrochemical imports from the Soviet Union, and the continuing US trade embargo.

Such small-scale and organic local production is now attracting the interest of some US researchers, who are keen to explore environmentally friendly alternatives to industrial farming. "I take my students to Cuba because of the contrast with US agricultural systems," says Catherine Badgley of the University of Michigan in Ann Arbor, who studies small-scale farming systems.

Badgley admires the way in which Cuban researchers are helping the country's growing number of small farmers develop new crop varieties and cheap methods of pest control that are free of synthetic chemicals. "People in Cuba are going into farming from other professions," she notes. "The opposite is happening in the rest of the world." Indeed, Franchi only took up agriculture when his construction business began to struggle in the early 1990s.

Franchi's farm and the gleaming labs of Havana's Scientific Pole seem worlds apart, but

Cuban scientists see a common theme. Neither would exist, they say, without socialist policies for applying science to local communities and setting research agendas in terms of public need. Franchi and his academic collaborators, for example, enjoy close ties with village mayors, and advise them on how to deal with everything from hurricane damage to energy efficiency. Cuba's biotechnology institutes, in turn, take advice on priorities from the country's extensive network of family doctors

For the people

Up in the hills of Las Terrazas, a 250-square-kilometre evergreen forest reserve that is home to more than a hundred bird species, the tight links between science and the government's social policies are particularly clear. When Castro came to power, the hillsides of Las Terrazas had been stripped practically bare for agriculture and fuel. Since then, forestry scientists and conservation biologists have overseen the planting of some 6 million trees — mostly indigenous species such as teak and mahogany. The residents of a new town built in the heart of the reserve were recruited to run the forestry projects and, since the 1990s, ecotourism schemes that bring in around 25,000 visitors a year. "Las Terrazas is a remarkable place because people live here and support the mission of the reserve," says Badgley. "There are terrible confrontations between conservationists and indigenous people in other parts of the developing world."

Visiting model initiatives such as Las Terrazas and the Scientific Pole, it's easy to think that Cuba is working miracles in science. But a trip to Roberto Cao Vazquez's chemistry lab at the University of Havana reveals the grim reality for researchers whose interests don't chime with the government's priorities. I enter on a steamy afternoon, having picked my way down the rubble-strewn passage outside. In

"I lost my position in the university and realized that my name had been erased from presentations I used to co-author." — Cuban expatriate

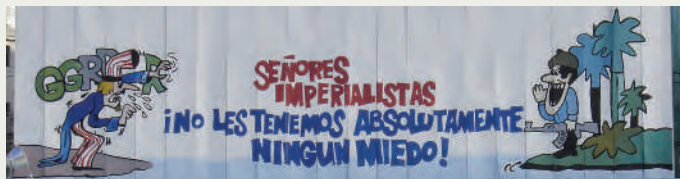
Neighbourhood dispute

Cuba has no commercial advertising, so billboards are rare in Havana. But several can be found facing the US Interests Section, the closest thing the United States has to an embassy in the capital. Each displays a strongly worded anti-US sentiment. The one pictured here translates as: "Imperialists, we have absolutely no fear of you."

Few subjects unite Cuban scientists as much as frustration with the policies of Cuba's superpower neighbour towards the island. A trade embargo, imposed in 1962 in response to the nationalization of US-owned businesses and Cuba's links with the Soviet Union, forces up the cost of scientific equipment. And most

scientists have given up applying for visas to visit the United States, saying they are tired of being refused by staff at the US Interests Section.

Slow Internet access is another consequence of the embargo. Cuban computer networks



cannot tap into the optical fibres that run past the island, because the connections are run by US companies. As a result, the University of Havana uses a satellite link. Because it is

expensive, the whole university is forced to share a connection that is about the same speed as the broadband links available to home users in developed nations.

Other US restrictions disrupt the flow of research funds into Cuba. Last year, for example, a team at the Center for Genetic Engineering and Biotechnology in Havana was awarded US\$700,000 by the Pediatric Dengue Vaccine Initiative. The initiative is based in South Korea, but funded in part by the US-based Bill and Melinda Gates Foundation — which must obtain clearance from the US government. Nearly a year after the grant was awarded, this permission has yet to be granted.

J. GILES/NATURE

the lab, the air-conditioning barely functions. A torn and yellowing periodic table hangs on the wall. The chemicals stores look pitiful. Cao, who works on fundamental aspects of supramolecular chemistry, is one of the have-nots of Cuban science.

Worlds apart

A jovial man with a pronounced American accent — the result of studying in a US-run school in Venezuela — Cao smiles wryly as he describes his plight. Journals: he doesn't have access to any, and relies on colleagues to send copies of interesting papers by e-mail. Chemicals: his budget is just US\$1,500 for the year. Equipment: his group has the country's only nuclear magnetic resonance machine, but it is so old that foreign researchers would laugh at it. "We're trying to do first-world science under third-world conditions," he says.

Cao and his colleagues survive by scraping together grants from abroad and through gifts of supplies brought by visiting foreign colleagues. As a result, they are able to publish in good journals, although progress in their labs is slow. "When I go and work with colleagues in Spain I am three times more efficient," says Cao.

Further evidence of Cuba's scientific divide comes from a visit to the University of Havana's marine biology lab. The team is respected by foreign conservationists, who admire the close links and influence the biologists have with Cuban leaders. But in the grand government scheme, studies of marine biodiversity rank well below vaccine development, and the scientists are often kept from their field sites because they can't afford fuel for their research boat.

"Our vessel is a Cuban innovation," adds group member Gaspar Gonzalez Sanson, to the laughter of his colleagues. "It's made of stone." Because carbon fibre is an expensive commodity in Cuba, the vessel's builders instead bent an iron-mesh frame into the shape of a hull and covered it with cement. It's

not pretty or easily manoeuvrable — but it works. "They're a wonderful group," says David Guggenheim, a marine biologist at Texas A&M University in Corpus Christi, who is working with Gonzalez to survey the biodiversity of Cuba's western shores. "But they're strapped for cash."

Scientists at Cuba's better-heeled institutes have little to say about the plight of researchers such as Cao and Gonzalez. And many Cubans are reluctant to discuss factors that hinder their work — such as rules governing foreign trips. Relative to most of Cuba's citizens, it is easy for scientists to travel. But if a researcher outstays the time permitted by the government, their status at home changes radically. "Then you can come back but only for short periods," says Cao. "It's a one-way ticket."

These restrictions present young researchers with a horrible dilemma. "Young people are very impressed when they come and see a big lab abroad," says one Cuban researcher working in Europe, who asked not to be named. But with many travel permits valid for just a few months, she explains, they then face a choice between sacrificing an opportunity to stay longer and do good research, or being separated from their families. A former professor from the University of Havana has first-hand experience of this, having overstayed his permission to work in Spain: "I lost my position in the university and realized that my name had been erased from presentations I used to co-author."

Given the hardships suffered by researchers outside the charmed circle of priority applied research projects, it is surprising not to hear more complaints from Cuban researchers.

"Cuba has become an experiment in scientific planning for countries that cannot afford to match the rich world's approach."

Government control may be one factor; open dissent is a risky policy in a non-democratic country. But equally important is an awareness that Cuba has battled against the odds to avoid the chaos and privations suffered by neighbouring countries such as Haiti. Older Cuban scientists, who remember the right-wing dictatorship that preceded Castro, are especially proud of what's been achieved.

The generally positive spin favoured by Cuban researchers is also reminiscent of first-world corporate culture. And to my surprise, many Cuban scientists and research managers are comfortable with this comparison — although quick to stress the differences. "The success is not sales, it's the impact on society," says Manuel Raíces Pérez-Casteñeda, a business development manager at the Center for Genetic Engineering and Biotechnology, the premier institute in the Scientific Pole. "We're chasing problems, not profits."

Setting an example

But can Cuba continue this chase in the longterm, having effectively turned its back on basic research? Halla Thorsteinsdóttir, a public-health expert at the University of Toronto in Canada, who has studied Cuban biotechnology, suspects that the lack of dedicated fundamental research may not be a huge problem. "It's hard to generalize across fields, but in biotech the boundaries between pure and applied research are fuzzy," she says. "And Cubans also have the expertise to take advantage of basic research done elsewhere."

Having so decisively concentrated its resources on a relatively small number of priority projects, Cuba has also become an experiment in scientific planning for countries that cannot afford to match the rich world's across-the-board approach. If it works, other nations are likely to go down the same route. "Developing countries can learn a lot from Cuba," Thorsteinsdóttir argues.

Jim Giles is a senior reporter for Nature, based in London.

BUSINESS

Pumping up the volume

The business of writing popular science books is hard to break into — and even harder to make money out of. Tony Reichhardt reports.

Skip Barker, one of a handful of US literary agents specializing in popular science books, has a sign up on his office wall saying “Yossarian Lives”. And the hero of *Catch-22*, Joseph Heller’s cult novel of the early 1960s, would be an apt patron saint for the book business: it’s hard to get a book contract unless you already have one, and publishers want unique ideas that are exactly like other unique ideas.

Yet Barker — whose Massachusetts-based Wilson Devereux agency does a steady trade selling his clients’ book proposals to science-friendly publishers — remains upbeat. There is still money to be made, and satisfaction to be had, from writing popular science books, even if they do not turn out to be blockbusters. “It’s all about the niche,” Barker says.

There is the occasional runaway bestseller, such as Stephen Hawking’s *A Brief History of Time* — which has sold more than 9 million copies in 40 languages since its 1988 publication. But many science books are considered successful if they manage 5,000 sales in hardback and the same again in paperback.

Hawking’s book, journalist Dava Sobel’s 1998 *Longitude*, and a handful of other hits have created the widespread impression of a boom market in science books. And the past few years have seen high-profile literary agents such as New York’s John Brockman negotiate six- and seven-figure contracts for celebrity scientist-writers such as Richard Dawkins (*The Selfish Gene*, *The Blind Watchmaker*) and Jared Diamond (*Guns, Germs, and Steel*).

But Barker points out that these hits have always been the exception. People in publishing say that the market for popular science books has cooled after a lively spell in the 1990s. It isn’t that the number of interested readers has dropped, says Peter Tallack, a London-based agent with Conville & Walsh who specializes in science books. “It’s that the number of titles has increased,” he explains, “so the readership is spread more thinly.” In Britain, however, data from Nielsen BookScan show revenues for one category — popular physics — falling from £3.6 million (US\$6.3 million) in 2001 to £2.2 million in 2004.

Shrinking budgets

Some publishers have cut the number of science titles they are printing, and marketing budgets are shrinking. Advances — the publishers’ upfront payments to authors — are down, too. Ian Stewart, a prolific author and mathematician at the University of Warwick, UK, whose 80 published titles include 20 or so written for the popular market, is a client of Tallack’s. Stewart says that non-celebrity authors do well to get a \$25,000 advance. Typical advances are half that.

Nor can you count on past trends to guarantee commercial success. A small boom in science-cum-history books modelled after *Longitude* — which told the story of the eighteenth-century quest for reliable maritime navigation — seems to have run its course. As Stewart puts it: “A little bit of sanity has

returned to the whole business.” But Stephen Morrow, the executive editor of New York-based Pi Press, predicts that science book publishing will continue to grow in the long run. “It may have dips and turns, but it won’t stop unless civilization does.”

Princeton University Press in New Jersey has been printing popular science books since Einstein’s *The Meaning of Relativity* in 1923. Most of the 60 titles it publishes each year are geared toward academic readers, with fewer than ten of the titles aimed at the lay reader, says editor-in-chief Sam Elworthy.

Spellbound: but unlike Dava Sobel’s *Longitude*, not all popular science books fly off the shelves

Princeton prefers authors who are real scientists to the science journalists responsible for many pop science titles. “Our usual model”, explains Elworthy, “is to go out and find people who are publishing top papers and say ‘How about writing a short book?’” The author’s reputation as an expert helps with marketing, he says. Among the scientists who have answered calls from Elworthy’s press in the past few years are Harvard astronomer and supernova expert Robert Kirshner (*The Extravagant Universe*) and Oxford geologist Simon Lamb (*Devil in the Mountain*). Good speakers are especially prized, as an assured performance on radio or television shows can boost sales.

The big pitch

Agents such as Barber work the deal in the other direction, pitching authors’ book ideas to publishers. Barber tends to work with journalists rather than scientists, and has been a steady supplier of authors to the Joseph Henry Press, a popular science imprint started by the Washington-based National Academies Press. The agents make a living by navigating the web of royalty deals and foreign publishing rights that remain a mystery to many authors.

Mathematician Stewart has both pitched ideas to publishers and been pitched to. With books such as *Flatland*, a re-imagining of the classic mathematical fantasy *Flatland*, and *What Does a Martian Look Like?*, coauthored by Jack Cohen, Stewart’s output is as varied as it is prolific.

Stewart is also one of the few academics whose job description includes writing for

NIelsen BOOKSCAN

TOP TEN SCIENCE BOOKS IN THE UNITED STATES (BEGINNING OF JULY 2005)

	Title	Author	Publisher
1	<i>A Short History of Nearly Everything</i>	Bill Bryson	Random
2	<i>Stiff: The Curious Lives of Human Cadavers</i>	Mary Roach	Norton
3	<i>The Fabric of the Cosmos</i>	Brian Greene	Random
4	<i>The Elegant Universe</i>	Brian Greene	Random
5	<i>Animals in Translation</i>	Temple Grandin	Simon
6	<i>A Brief History of Time</i>	Stephen Hawking	Random
7	<i>The Road to Reality</i>	Roger Penrose	Random
8	<i>Chemistry for Dummies</i>	John T. Moore	Wiley
9	<i>The Cartoon Guide to Chemistry</i>	Larry Gonick	Harper
10	<i>The Genius Factory</i>	David Plotz	Random

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

BRAND X PICTURES/ALAMY

popular audiences. Most researchers struggle to find the time. For those at the top of their fields, says Elworthy, "book writing usually takes a back seat". That's why Stanford neuro-endocrinologist and stress expert Robert Sapolsky (*A Primate's Memoir*) never takes assignments with tight deadlines, and refuses to do out-of-town publicity tours. He already has a demanding career, he says. Popular writing is "meant to be fun".

Not that you can't profit from it. Stewart says he generally gets his payment upfront, and doesn't worry as much about downstream royalties. The advance should therefore be large enough to constitute fair remuneration but not so large as to risk ostracism from the publisher if the book doesn't sell well. "Sometimes it's best to build up slowly," advises Tallack.

Aspirant authors have two main choices to make: the subject matter, and the level at which the book should be pitched. A snapshot of top-selling science books for one week in July (see Table, left) includes such lofty tomes as Brian Greene's *The Elegant Universe*, but also *The Cartoon Guide to Chemistry*. Other possible approaches include books tied to other books or movies, and children's science books. According to one survey, the surprisingly healthy market for the latter accounted for 12% of all juvenile book sales in Britain in 2003, outselling books on sports.

When it comes to subject matter, most scientists stick to their own field of research — and some are more popular than others. Astronomy and physics, for example, sell well, Barber says. But with bestsellers notoriously difficult to predict, most professionals offer would-be authors the same, not particularly helpful advice: tell a good story in an engaging, witty style. "Write about what you think is important," suggests Morrow. And above all, make your book different — but not too different — from what's already filling the shelves and the bargain bins. ■

IN BRIEF

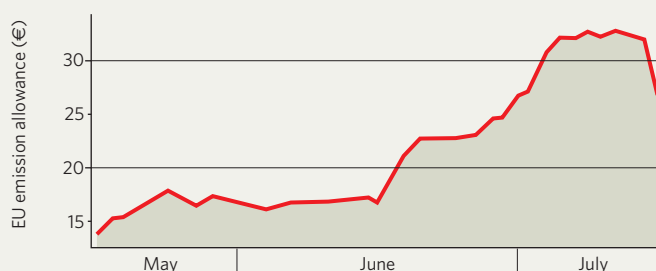
SPEEDY VACCINE US regulators have granted the maker of a promising influenza vaccine 'fast track' status in its licensing application — increasing the chances that it will be ready in time for the flu season after next. ID Biomedical of Vancouver, British Columbia, already supplies the injectable vaccine, called Fluviral, in Canada. US distributors have committed to making 38 million doses of it available each year if the vaccine wins approval from the US Food and Drug Administration.

CHIP WARS European Commission officials launched a series of dawn raids on offices of Intel, as part of an investigation into allegations that the chip-maker has been abusing its dominant position in the market for microprocessors in personal computers. The aggressive action by the commission's competition directorate on 12 July led some analysts to conclude that it thinks it has a strong case against California-based Intel. Rival chip-maker AMD has filed complaints to the directorate about Intel's business practices twice, in 2000 and 2004.

BIOTECH LEADER SOARS Stock in Genentech rocketed to almost \$90 after forecasts of increased earnings and promising results from a trial of its drug Tarceva in lung cancer patients. The San Francisco-based biotechnology company's healthy second-quarter results were based on stronger-than-expected sales for Tarceva and other cancer drugs. Genentech's market capitalization has more than doubled since March to some \$90 billion — exceeding that of major pharmaceutical companies, such as New Jersey-based Merck.

MARKET WATCH

EUROPEAN CARBON INDEX



High natural-gas prices and a summer heatwave are pushing the cost of European allowances to emit carbon dioxide sharply upwards.

At the European Energy Exchange (EEX) in Leipzig, Germany, an allowance entitling the bearer to emit one extra tonne of carbon dioxide per year peaked earlier this month at almost €30 (US\$36) — double its value in May. The price subsequently dipped somewhat (see Graph).

Since January, large companies in Europe have been required to possess an allowance for each tonne of carbon dioxide they emit. If they need more allowances than they get from their national government, they can purchase additional ones on the market.

The EEX launched its European Carbon Index last October, based on daily notifications from the trade

departments of eight large energy companies. The value of the options has soared from about €8 since trading started in January.

Stefan Kleeberg of Frankfurt-based Climate Change Consulting explains that the high price of natural gas is boosting the generation of electricity from coal — as is a warm, dry summer in southern Europe, which is cutting generation from hydropower and from some coolant-thirsty nuclear power plants. Higher carbon dioxide emissions from coal burning may be fuelling demand for extra allowances, he says.

But trading volume remains small, and no one is sure whether the current high prices are sustainable — or just a quirk caused by the natural volatility of a young market. Kleeberg thinks demand for the allowances will stay strong. "The sentiment is bullish," he says.

Quirin Schiermeier

EUROPEAN ENERGY EXCHANGE

Unlike climate science, GM is full of uncertainties

SIR — Greenpeace has been invited by David Dennis, in Correspondence, to reconsider our opposition to genetically modified (GM) crops in the light of our support for the consensus on climate science (“Activists should accept mainstream view of GM” *Nature* **435**, 561; 2005). There are two factors to consider when deciding to support any apparent scientific consensus.

First, what was the process of arriving at the current mainstream position? In the case of climate change, uncertainties over the physics, measurements, modelling and historical data have generally (although sometimes erratically) tended to be resolved. In the case of GM, further investigation of genomes and gene function has led to new insights, such as alternative splicing mechanisms and the regulatory roles played by RNAi and chromatin packing, which question the fundamental understanding of gene regulation and control. This is demonstrated by the hedging on certainties in the UK government’s GM science panel review in 2003, which was a far cry from the certainties expressed in the mid-1990s.

Second, as an environmental-protection organization, it should come as no surprise that we interpret scientific uncertainty in favour of environmental protection. Anyone who interprets a given level of uncertainty to propose a policy action must be imposing their values, even if that means advocating “do nothing now except more research”.

Whereas we are clear about our values, one might wonder what values are espoused by those, especially in the United States, who support GM organisms but reject the evidence of climate change.

Douglas Parr

Greenpeace UK, Canonbury Villas,
London N1 2PN, UK

Leave GM analysis to the relevant scientists

SIR — David Dennis, in Correspondence, claims that “an overwhelming majority of plant geneticists, biochemists and molecular biologists have endorsed the use and safety” of genetically modified (GM) crops (“Activists should accept mainstream view of GM” *Nature* **435**, 561; 2005). I question the validity of that claim.

Assessing the potential environmental and/or economical consequences of using GM crops — such as their impact on soil fauna or on non-target organisms — requires analysis in crop fields and in the natural environment, working on relevant objects, at the relevant scale.

As questions about the use and safety of GM crops concern primarily environmental science, statements by biochemists and molecular biologists, who deal with simplified biological systems, at small scales, only add to the problem of misinformation and lead to an increase in concern about GM crops.

“Statements by scientists who deal with simplified biological systems, at small scales, only add to the problem of misinformation.”

— Denis Couvet

Perhaps the public would be less worried if it was the overwhelming majority of environmental scientists who felt confident about the use and safety of GM crops.

Denis Couvet

Department of Ecology,
Muséum National d’Histoire Naturelle,
55 rue Buffon, 75005 Paris, France

Compensation for climate change must meet needs

SIR — Sujatha Byravan and Sudhir Chella Rajan, in Correspondence (“Immigration could ease climate-change impact” *Nature* **434**, 435; 2005), argue that major greenhouse-gas emitters should provide compensation for the impacts of climate change.

We believe that compensation is on the cards. We also believe there will eventually be the science to accurately establish liability for those impacts (see M. R. Allen and R. Lord, *Nature* **432**, 551–552; 2004). But the nature of compensation will always remain contested. The notion that migration is a sustainable adaptation strategy for future climate change ignores the fact that patterns of migration are strongly based on social networks and cultural links. Legislating for flows of people may simply not appeal to migrants.

New Zealand’s creation of the Pacific Access Category, in response to concerns about climate change, is an instructive example. The scheme allows for up to 75 people from Tuvalu to migrate each year, but since it began in July 2002, fewer than half the places available have been filled. This possibly suggests that even in Tuvalu, where there is widespread concern about climate change, people are not eager to leave their homeland. This example points to the need for policies and measures that help people adapt to climate change, in order to lead the kind of lives they value in the places where they belong, rather than to encourage migration.

The United Nations Framework Convention on Climate Change (UNFCCC) is developing mechanisms to help vulnerable countries adapt to climate change. However,

these mechanisms are not the obligatory compensation transfers that Byravan and Rajan espouse. Indeed, the prospect of enforceable migration threatens voluntary processes, such as the UNFCCC, as they may deter large greenhouse-gas polluters from participating in the process.

We believe that promoting and funding activities that enhance *in situ* adaptation for vulnerable populations is a more practicable and equitable approach than migration-based compensation strategies.

W. Neil Adger*, Jon Barnett†

*Tyndall Centre for Climate Change Research,
University of East Anglia, Norwich NR4 7TJ, UK

†School of Anthropology, Geography and
Environmental Studies, University of Melbourne,
Melbourne, Victoria 3010, Australia

There’s more to a colourful life than simply sex

SIR — Rolf Hoekstra in News and Views (“Why sex is good” *Nature* **434**, 571; 2005) surely overstates the contribution of sexual reproduction to the aesthetic appeal of nature. He assumes that complex organisms similar to flowering plants, insects and peacocks would evolve, but that without sex the world would be drab and colourless. However, asexually reproducing plants would probably still take advantage of animals for spore dispersal, producing fruits or other rewards (advertised by special structures) as ‘payment’. Indeed it is likely that our colour vision developed in order to detect ripe fruits, which have little to do with sex and a lot to do with seed dispersal.

Similarly, although it is true that many gaudy avian displays are aimed at mate attraction, there are many brightly coloured animals whose decoration serves as a threat or warning. Colonial bees, wasps and many butterflies and moths all bear distinctive and colourful markings discouraging interference.

Sexual reproduction, in maintaining genetic heterogeneity within a population, is clearly a major mechanism by which species survive catastrophes and adapt to the subsequent conditions. However, the pro-sex lobby always seems to downplay the importance of asexual reproduction to evolution in stable environments. Apomictic organisms (not strictly asexual of course), such as dandelions, can develop localized populations accumulating mutations that render them distinct from other members of their clade. This kind of diversification is rarely seen in sexually reproducing populations unless they are subjected to selective pressure.

Sex is good, but it ain’t everything.

Paul Kenton

University of Wales, Aberystwyth SY23 3DA, UK

BOOKS & ARTS

Letters from a hero

What made Richard Feynman so much more than a Nobel prizewinning physicist?

Perfectly Reasonable Deviations from the Beaten Track: The Letters of Richard P. Feynman

edited by Michelle Feynman

Basic Books: 2005. 512 pp. \$26

Published by Allen Lane in the UK as *Don't You Have Time to Think?* £18.99

Peter Galison

Richard Feynman was a physicist's physicist. To have been a principal builder of quantum electrodynamics (QED) — joining special relativity and quantum mechanics — was accomplishment enough. But his youthful version of QED was more than that: it became a model solution for quantum field theories more generally. His eponymous diagrams became for theorists what a hammer and nails are for carpenters. In later years he struck gold again. His 'parton' model was not only helpful in classifying elementary particles, it also made quarks seem more real as they scattered electrons when these hit them inside protons and neutrons. His account of superfluidity in liquid helium and his model of weak interactions also had a lasting impact.

Then there were his interventions in the world beyond fundamental research. At Los Alamos, Feynman — then in his mid-20s — contributed to our understanding of fission in the core of a nuclear bomb. That work earned him enormous respect from his colleagues, especially Hans Bethe, who led the theory group at Los Alamos.

Much later in life, Feynman very publicly intervened in the analysis of the 1986 Challenger space shuttle disaster, famously dunking an O-ring into freezing water to show the disastrous effects of winter weather on this crucial part. Suddenly, it was all too clear that O-rings had failed to remain flexible enough to seal hot gases inside the booster.

Without these fundamental contributions, Feynman would not have been Feynman. But his fame within and outside the physics community exceeds greatly that of other Nobel prizewinners. Even Bethe — who won the 1967 Nobel Prize in Physics for figuring out why the Sun shines, and who powerfully opposed the arms race — does not have the hero status of Feynman. Young physicists regularly tack a poster of Feynman above their desks. If there are posters of other Nobel prizewinners on sale, I haven't seen them.

Except, of course, for Albert Einstein.



M. FEYNMAN AND C. FEYNMAN

Pin-up: young physicists often tack a poster of Feynman above their desks for inspiration.

Although Einstein's iconic status extends far beyond the physics world in ways that Feynman's does not, there are similarities in the way their public personae have been mythologized. To a certain degree, Einstein has been reinvented by each generation, but he has come to stand for the isolated genius, the outsider who helped to propel the century into modernity. Just about every modern poet wrote about Einstein. Artists, architects and musicians used him as a muse. Many still do, a century after his miracle year of 1905. And yet, since the early 1960s, generations of science students have held Feynman, not Einstein, as their model and guiding star.

Why? Perhaps because Feynman has come to stand for a kind of rough-edged counterweight of reason to tradition and pomposity. Here was a truth-teller with a Brooklyn accent, an unaffected clear thinker who would, in the rigor of his intuitions and calculations, be unmoved by social niceties. Although Einstein spent more than two decades in the United States, often plunging himself into the politics of disarmament and anti-McCarthyism, he never lost his European identity. Feynman, by contrast, took pride in his Runyonesque Americanness, in his pragmatism, in his indifference to title, scholarly learning and unearned awards.

In a letter rejecting an honorary degree, Feynman wrote: "I remember the work I did to

get a real degree at Princeton and the guys on the same platform receiving honorary degrees without work — and felt an 'honorary degree' was a debasement of a 'degree which confirms certain work that has been accomplished.'" He swore then and there he would not accept such undeserved celebration were he to be offered an honorary degree again.

Michelle Feynman, Richard's daughter, has edited a remarkable set of letters (including the above) that provide us with a much clearer view of the non-physicist Feynman. This is not the Feynman analysing liquid helium, but the Feynman who grappled with his increasing fame, his relation to other physicists, his family and, most movingly, with his first wife, Arline, who died of tuberculosis in June 1945.

The letters to his wife, clear and loving, culminate in one written after her death: "It is such a terribly long time since I last wrote you — almost two years but I know you'll excuse me because you understand how I am, stubborn and realistic; and I thought there was no sense to writing... I'll bet that you are surprised that I don't even have a girlfriend... after two years... You only are left to me. You are real. My darling wife, I do adore you. I love my wife. My wife is dead. Rich. P.S. Please excuse my not mailing this — but I don't know your new address."

Eventually Feynman remarried, had children, enjoyed people and ideas, and most of all

delighted in physics. But there is something a bit muted in the later letters, especially the more intimate ones. Much of Feynman's passionate engagement flowed through his physics. This comes across remarkably well in the letters Michelle Feynman has chosen — in his enthusiastic responses to friends, colleagues, students and physics hobbyists. He advised, cajoled and encouraged: Try this, think about that, have courage in your ideas, think for yourself.

Occasionally, Feynman's passion for physics — for control over a world he imagined he could create entirely by himself — slid into a disdain toward everything non-scientific. In one exchange, he pronounced dismissively on poetry and in particular on poets' insufficient appreciation of physicists' vision of the world: "My lament was that a kind of intense beauty that I see given to me by science, is seen by so few others; by few poets and therefore, by even fewer more ordinary people."

Although Feynman's physics at times resem-

bles the physics of the young Einstein, his presence in the world is very different. Einstein never lost his fascination for philosophy, for Kant or for his near-contemporary, Poincaré. Feynman found philosophers nothing but a burden, a vulture-like presence that swooped in when strong ideas were dying. And while Einstein came to believe that physical reality lay deep in mathematical physics, Feynman never gave up hoping for a physics driven, at bottom, by an almost tactile intuition.

Much of Einstein's life found him cast and self-cast as an oracle. Feynman preferred the persona of a fast-draw street-smart kid. Yet beyond these striking differences, both Einstein and Feynman found ways to hold their own, fiercely maintaining their positions as individuals in a time when physics and fame, as never before, pressed them to assume their place in teams and groups. ■

Peter Galison is in the Department of Physics, Harvard University, Science Centre 235, 1 Oxford Street, Cambridge, Massachusetts 02138, USA.

Principles of Geology. Above all, he had the stimulus of intelligent discussion at the Geological Society in London after his return.

After spreading his attention widely at first, he concentrated increasingly on one of geology's focal problems at the time: that of crustal elevation and subsidence. This underlay his interest in the effects of the great earthquake he witnessed in Chile, as well as his fieldwork in the high Andes. It led him to out-Lyell Lyell — giving an even better explanation in terms of observable processes — with his innovative theory of coral reefs, sketched in outline before he ever saw one, in which corals functioned simply as markers of crustal movement.

With sublime confidence, he anticipated that the "geology of whole world will turn out simple". His 'theory of the Earth', like Lyell's, was one that envisaged the ceaseless movement of crustal plates, not horizontally as in modern plate-tectonic theory, but vertically. However, after his return to Britain he cited the famous Parallel Roads of Glen Roy in Scotland in support of his theory, only to be upstaged by the Swiss naturalist Louis Agassiz's glacial theory, which was a far better explanation for the puzzling terraces. This was a critical challenge to 'simplicity' in the light of which Darwin eventually conceded that his own effort had been a "gigantic blunder".

Interspersed with Herbert's valuable analyses of Darwin's geological fieldwork and theorizing are chapters on other topics. In line with current trends in the historiography of other sciences, she describes in fascinating detail the practical aspects of Darwin's geology: his hammer and other instruments, his methods for collecting specimens and making notes, and so on. She also discusses the Romanticism of the travel narratives that he took as his literary models, and the contemporary debates in England about geology and Genesis. And, perhaps of greatest interest to other Darwin scholars and to biologists, she analyses with care the ways in which his geology generated the problems to which his eventual theory of the origin of new species was designed to be the solution.

I have only two reservations about this fine volume. The first is that Herbert tends to underestimate the extent to which Darwin was developing lines of research already being explored by other geologists — not only his hero Lyell but also those who were critical of Lyell's theories. Second, like Darwin himself — who was only fluent in English — she does not adequately emphasize the thoroughly international character of the geological world during the most creative period of his life. Nonetheless, this is a highly important contribution, not just to Darwin studies but also to the sadly neglected field of the history of geology itself. ■

Martin Rudwick is in the Department of History and Philosophy of Science, University of Cambridge, UK. His latest book, *Bursting the Limits of Time*, will be published by the University of Chicago Press in August.

Darwin's first love

Charles Darwin, Geologist

by Sandra Herbert

Cornell University Press: 2005. 512 pp.
£21.95, \$39.95

Martin Rudwick

In 1836, Charles Darwin returned to England after his five-year voyage on *HMS Beagle*. He soon became a closet evolutionist, working on his biological theory of evolution before publishing *On the Origin of Species* in 1859. These bare facts are not incorrect, but they are seriously incomplete. Two years after his return, reflecting on his life so far, he described himself as "I the geologist..." This was not an isolated remark — it expressed his chosen identity. Indeed, it was as a competent geologist that Darwin first came to the attention of the scientific community and made his name as a promising young 'man of science'.

His main interests gradually shifted sideways from geology into zoology and botany, but it is deeply misleading to read his career as if there was an inevitability about the move. This has been well known to Darwin specialists, but Sandra Herbert's *Charles Darwin, Geologist* is the first full-length treatment of his geological research, describing and analysing the work in its own right as well as in its role as a foundation for his later biological work.

Herbert's approach is not strictly biographical; some background knowledge of Darwin's life — best gleaned from Janet Browne's superbly readable two-volume *Charles Darwin* (Jonathan Cape, 2003) — is, in effect, taken for granted. Herbert treats Darwin's geological work as a series of specific topics, with a chronological analysis of each. Although she

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Blank slate: there was nothing inevitable about Darwin's evolution from geologist to biologist.

covers in outline almost his entire career, she focuses on Darwin's Beagle years, his fruitful few years in London, and the first years of his long life at Downe in Kent — in other words, on the 1830s and early 1840s.

As a highly respected member of the scholarly 'Darwin industry', she has an enviably thorough knowledge of the vast Darwin manuscript archive. She makes full use of the revealing details of Darwin's famous scientific notebooks and his voluminous correspondence, which greatly deepen our understanding of his published work.

Herbert rightly emphasizes that the geology to which the young Darwin contributed was already a well-established science. He had excellent informal training from Adam Sedgwick and John Henslow at the University of Cambridge, support from a substantial scientific library on board the Beagle, and inspiration from Charles Lyell's newly published

DOCUMENTARY

In the right place at the right time

Guns, Germs and Steel

directed by Tim Lambert
National Geographic Television & Film,
premiered on PBS in the United States on
11, 18 and 25 July.
Also available on DVD in North America.

Henry Gee

One day on a trip I took to Kenya in 1998, the cook was suffering a bout of malaria so we had to fend for ourselves. The local Turkana were immune to the disease, but nearly all of those from elsewhere — including the cook, from the malaria-free highlands around Nairobi — were susceptible. The cook was right as rain the next day, but millions of Africans are less fortunate.

In some parts of the continent, malaria remains the biggest killer of children under five. To be moved by a bald statistic such as this, we have to be confronted with the reality on the ground, as polymath and prizewinning author Jared Diamond found when he visited a hospital in Zambia — where children were dying all around him.

“There’s a difference between understanding something intellectually and experiencing it firsthand,” says Diamond, who was filming a TV adaptation of his book *Guns, Germs and Steel* (Norton, 1997). “In my book, ‘germs’ was one of the three main forces of history, and it’s impersonal, and it’s still a different entity... it hits me to be in a place where germs are in action.”

Diamond’s ‘guns, germs and steel’ theory is economic geography at its grandest, and like many such ideas, emerged from a small event. Diamond has been a regular visitor to Papua New Guinea since the 1960s, where he watches the birds and, by extension, the people. He has found that despite their stone-age technology, New Guineans are at least as resourceful as Westerners. So why is it that the world is dominated by Europeans, rather than another group? It took a question from a New Guinean to coalesce the problem in Diamond’s mind: “Why is it that you Westerners have so much cargo, when we New Guineans have so little?” It took 30 years for Diamond to formulate his response: *Guns, Germs and Steel*.

According to Diamond, cultures succeed not by the inherent cleverness of their people, but by the luck of geography and biology. Western civilization started 10,000 years ago in the Middle East, which happened to host the wild plants and animals that were the easiest and most productive to domesticate. Based on wheat, cattle and sheep, this civilization spread throughout the temperate regions of Eurasia. The other civilizations it met were isolated by geography, so they could not spread, and had fewer domesticable species — in the case of the



On target: Jared Diamond believes accidents of geography determine a society’s progress.

farmers of the New Guinea highlands, so few that the surplus needed to support non-farming members of society could never be achieved. For them, technology would always be out of reach, as would resistance to diseases such as smallpox that owe their origins to long association with domestic animals.

This is why tiny groups of sword-wielding, gun-toting, horse-riding, smallpox-carrying but otherwise unremarkable Spanish adventurers wiped out the Inca and Aztec civilizations within a few years; how a few Boers with muzzle-loading rifles destroyed the mighty Zulus and Matabele in less time — in the great sweep of history — than it takes to make a cup of bush tea; and why you are reading this article in English rather than Nahuatl.

But whenever European settlers in southern Africa strayed north, their advantages gave out. European-style crops would not grow in the less temperate climate, and they and their animals wasted away from tropical diseases. The Bantu farmers they met were unaffected, having built their success in a durable but diffuse civilization that stayed away from rivers or large population centres, and had thus reached an understanding, if not an equilibrium, with disease. The colonial imposition of railroads and large cities eventually destroyed all that, which is why modern Africans now suffer from the ills that were once much less of

a problem, and why — in the final part of this three-part series — we see Diamond break down in a hospital in Zambia.

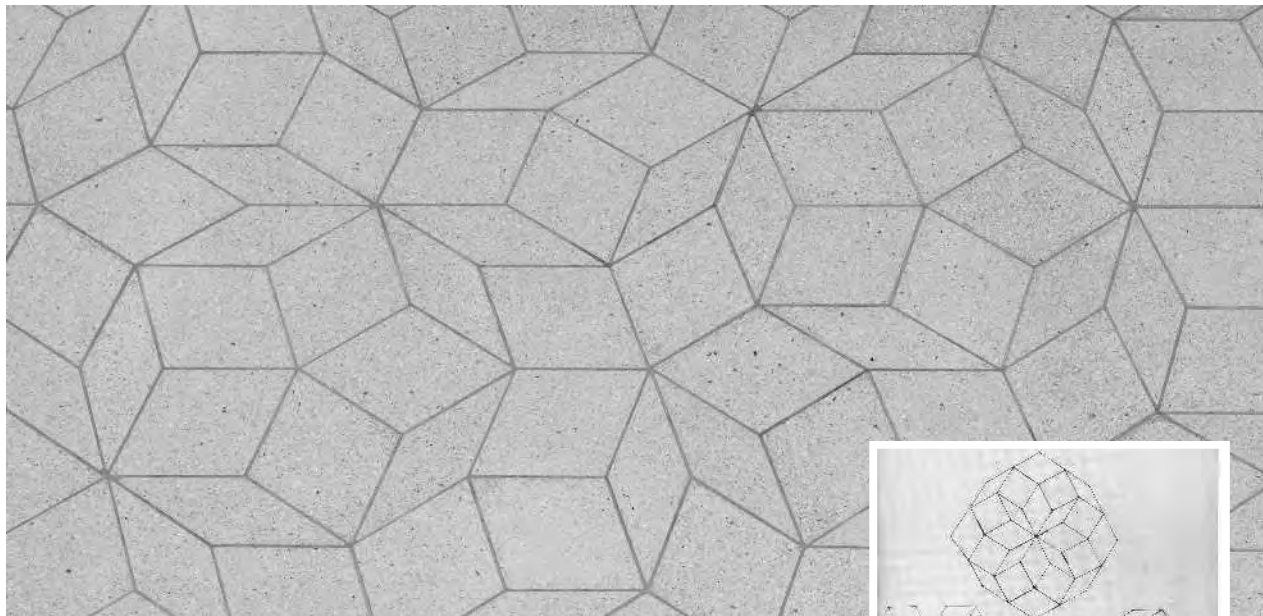
Not everyone will agree with Diamond’s thesis. It stems from a new view of history that condemns concepts such as ‘manifest destiny’ as narrow or racist, and is prepared to examine ‘counterfactuals’ — scenarios of what might have happened, had the starting conditions been different. But if the new school of history teaches us a lesson, it is that once one understands the reasons for one’s predicament, they can be transcended. Singapore and other tropical, southern Asian countries prone to disease, have overcome these obstacles to become big players in twenty-first century technology. Perhaps Africa might follow.

The three hour-long programmes are as beautifully shot as you would expect from National Geographic, but the first two and a half hours are too leisurely, making the final half-hour far too hasty. And although the actor Peter Coyote does a fine job as narrator, having Diamond play second fiddle in his own orchestra prevents *Guns, Germs and Steel* from achieving the magisterial air of similar programmes that have become classics. Would Alastair Cooke’s *America* have been half as good had Dan Rather narrated it, with Cooke himself in a walk-on part?

Henry Gee is a senior editor at *Nature*

A trick of the tiles

Penrose tiling is realized on a huge scale in Perth to give a perceptual feast for the eyes.



TERRACE PHOTOGRAPHERS

Martin Kemp

Geometry in Western art predominantly involves space and proportion. But in other cultures, most notably Islamic, Chinese and Japanese, artistic geometry flowered most conspicuously in flat patterns, above all in the invention of striking tessellations in tiling, mosaics and textile designs. The trick was to invent a repeated geometric pattern of considerable complexity without generating 'gaps' that had to be arbitrarily filled.

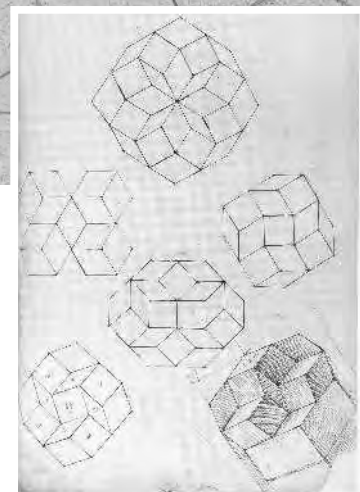
Working with straight-edge and compass, over the centuries geometer-artisans accumulated an astonishing array of periodic patterns, including many of the types later defined in mathematical theory. The basic principles involved in such tessellations — that they are possible only with one-, two-, three-, four- or six-fold symmetries — held good until 1974, when the mathematician Roger Penrose from the University of Oxford unveiled a 'gap-free' tiling based on five-fold symmetry. The two types of tile are the 'kite' and the 'dart', obtained by intersecting a single rhombus with angles of 72° and 144° , respectively. As it happens, the ratio of the long edges to the shorter is the golden ratio.

Later, Penrose demonstrated that a simpler pair of tiles in the form of fat and thin rhombuses (or diamonds) also generate five-fold symmetry in an aperiodic pattern. In other words, from the five equivalent directions from each angular point, the pattern does not repeat itself — though instinctively we feel that it should if we look hard enough.

This more recent Penrose tiling has been cleverly used in the floor of the atrium of the new Molecular and Chemical Sciences Building at the University of Western Australia in Perth. The floor (see above) was the idea of Professor David Kepert, then head of the School of Chemistry, and his colleague Frank Lincoln, and the tiling was developed by the architect Gus Ferguson. Starting from a central five-pointed star midway between the lift and the stairway opposite, Ferguson used two types of locally manufactured concrete tile in the form of fat and thin rhombuses to develop the five-fold symmetry across the entire floor.

The result is a strange spider's web of lines, which, although composed of just two simple repeated shapes, presents us with a perceptual and cognitive field replete with possibilities. As with the patterns designed over the centuries, its fascination transcends the purely mathematical. Looking at such arrays, we are perceptually tuned to tease out coherences that go beyond seeing the paired shapes.

When staring at the patterns for any length of time, a deep-seated perceptual proclivity comes into play. We can hardly avoid discerning shapes compounded from clustered elements, discovering five-pointed stars and bilaterally symmetrical polyhedra, plotting zigzag strands and so on. Almost inevitably, spatial instincts also come into play, although none of the 'sides' of the perceived 'bodies' converges to notional vanishing points. We can, for instance, play Necker cube-type games



with apparent octagons, and facet the surface into a kind of cubist medley of receding and advancing planes.

Authors of earlier tiling patterns deliberately enhanced the implicit spatial thrusts by infilling the tiles with varied colours and tones in regular repeats, and by alternately interweaving the thin bands that demarcate the patterns. This effect is seen clearly in the 1524 doodling of the German artist Albrecht Dürer, one of whose repeated patterns of flat rhombuses and cubes was thrown suddenly into paradoxical relief through the addition of hatched shading (see above, inset). We instinctively tend to do the same in our minds, even where the pattern frustrates perspectival coherence.

It is easy to imagine why the designer of the Perth floor did not shade or colour the tiles. The spatial cacophony that would result with an aperiodic pattern on such a scale would have been jarring. But the waiting visitor or someone pausing on one of the landings can easily play the spatial game. Indeed it is hard to avoid doing so. Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK, and is the author of *Leonardo* (Oxford University Press, 2004).

PALAEOCLIMATE

Foreshadowing the glacial era

Lee R. Kump

Under what circumstances do glaciations persist or occur only transiently? Indications that short-lived 'icehouse' conditions occurred during the otherwise warm Eocene provide further cause for debate on the question.

Earth entered its present glacial state 34 million years ago with the growth of the Antarctic ice sheet¹. This major climate transition occurred abruptly and essentially irreversibly at the Eocene–Oligocene boundary, a conclusion based on the record of ice-sheet size preserved in the oxygen isotopic composition of limestones^{2,3}. The preceding Eocene epoch (55–34 million years ago) is generally considered to have been warm and ice-free, but data on this time interval, as recorded in cores of marine sediments, have been sparse.

By analysing newly acquired core material from the tropical Pacific, Tripati *et al.*⁴ (page 341 of this issue) provide a much more detailed view of the climate system before the permanent transition to the glacial state. What they find there, and in contemporaneous cores from the South Atlantic, is several small glaciations and one major (but transient) glaciation in the middle to late Eocene, millions of years before the Eocene–Oligocene boundary.

The water from which continental ice sheets grow derives from evaporation of ocean water and its deposition at high latitudes as snow. Thus, as ice sheets grow, sea level falls. Moreover, compared with sea water, the snow is enriched in the lighter isotope of oxygen, ^{16}O . So, as ice sheets grow, the ratio of ^{18}O to ^{16}O in the oceans increases; the ratio is generally presented as the standardized ratio $\delta^{18}\text{O}$. Organisms that precipitate skeletons of calcium carbonate (CaCO_3) do so close to oxygen isotopic equilibrium with the waters in which they grow, so the $\delta^{18}\text{O}$ of fossil skeletons provides a proxy measure for ice-sheet size in the past. However, as the equilibrium $\delta^{18}\text{O}$ of the CaCO_3 also depends on temperature, an unambiguous interpretation of ice-sheet size from fossil $\delta^{18}\text{O}$ requires additional temperature information. Tripati *et al.*⁴ use an independent temperature proxy — the amount of magnesium incorporated into CaCO_3 shells — to isolate the effects of changing ice volume.

The temperature proxy indicates that there was little global cooling associated with the late Eocene glaciations, suggesting that, as in

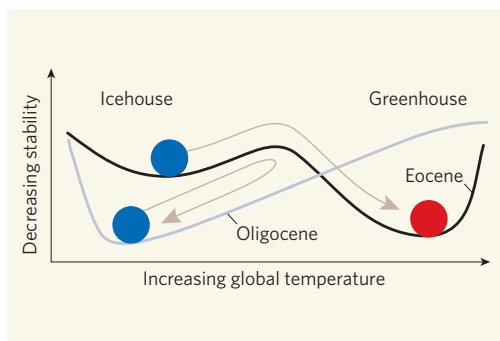


Figure 1 | Glacial stability and instability. Global temperature is indicated by the balls. The findings of Tripati *et al.*⁴ suggest that a glacial ('icehouse') climate state could have existed in the otherwise warm Eocene. But the 'greenhouse' state was perhaps readily re-established through instabilities arising from stochastic or orbital fluctuations (indicated by the arrow). As levels of atmospheric CO_2 fell, the stability of the icehouse state might have increased and that of the greenhouse state might have decreased, so that by the early Oligocene any such fluctuations would not have moved the system back to the greenhouse state, but would have been followed by rapid coolings to the stable glacial state. (After Fig. 2-3 in ref. 10.)

the earliest Oligocene event⁵, most of the shift in $\delta^{18}\text{O}$ was due to an increase in ice volume and that cooling may have been limited to high latitudes (or that ice-sheet accumulation there was limited by moisture rather than temperature). As in the earliest Oligocene, the isotopic data seem to require the presence of ice sheets on Antarctica at least as thick as those today, and substantial ice sheets in North America (most likely Greenland). This latter result runs contrary to conventional wisdom, which holds that the Northern Hemisphere glaciation began tens of millions of years later⁶.

From their analyses of accumulation patterns of CaCO_3 on the sea floor, Tripati *et al.*⁴ find signs of substantial perturbations in the ocean's carbon cycle during the Eocene, patterns that mimic those of the more permanent change to come. The oceanic calcium-carbonate compensation depth (CCD), the depth below which CaCO_3 does not accumulate because deep waters are corrosive, increased significantly during the glacial events. Deepening of the CCD is an expected

consequence of sea-level fall, because it allows for additional deep-sea carbonate accumulation that compensates for the loss of carbonate deposition from shallow waters⁷. So these data corroborate the claim that substantial ice sheets existed in the Eocene.

Why the Oligocene Antarctic ice sheet persisted but the Eocene ice sheets did not is unclear. As a driver for glaciation, Tripati *et al.* invoke a reduction in the amount of atmospheric CO_2 accompanying the growth of the Himalayas and resulting from enhanced chemical weathering of the rocks unearthed^{8,9}. They suggest that increased biological productivity in the early Oligocene, and so increased use of CO_2 in photosynthesis, may have provided the additional drawdown of atmospheric CO_2 that was necessary to sustain the glaciation. Perhaps the stability of the glacial state increased as atmospheric CO_2 levels fell, so that stochastic effects (such as volcanic eruptions releasing CO_2 , or destabilization of methane embedded in the sea floor), or variations in Earth's orbit, became insufficient to jar Earth out of its glacial state (Fig. 1). Future work on proxy measures of atmospheric CO_2 from the Eocene and Oligocene should provide the necessary test for this hypothesis.

The equatorial Pacific sediments analysed by Tripati *et al.*⁴ are thought to represent oceanographic conditions over a broad region of the Pacific, and the data from the South Atlantic support the proposition that these changes were indeed globally significant. Nevertheless, a general acceptance that glaciations occurred in the middle to late Eocene will probably require further evidence. The suggested existence of large Northern Hemisphere ice sheets in the Eocene is highly controversial. Moreover, the fidelity of the magnesium content of CaCO_3 as a measure of temperature demands further scrutiny. However, the existence of precursor glaciations foreshadowing the major transition to the glacial state is theoretically expected of a system that is subject to natural fluctuations but is gradually evolving from one stable state to another.

If decreasing atmospheric CO₂ stabilized the glacial state in the Oligocene, might increasing atmospheric CO₂ from fossil-fuel burning destabilize it in the future? The lesson to be learned here is that we should watch for subtle signs that we are moving from the icehouse world in which Earth has remained for 34 million years into a new, greenhouse world. ■

Lee R. Kump is in the Department of Geosciences, Pennsylvania State University, 535 Deike Building, University Park, Pennsylvania 16802, USA.
e-mail: lkump@psu.edu

1. Wise, S. W., Breza, J. R., Harwood, D. M. & Wei, W. in *Controversies in Modern Geology* (eds Mueller, D., McKenzie, J. & Weissert, H.) 133–177 (Academic, San Diego, 1991).
2. Kennett, J. & Shackleton, N. J. *Nature* **260**, 513–515 (1976).
3. Zachos, J. C., Quinn, T. M. & Salamy, K. A. *Paleoceanography* **11**, 251–266 (1996).
4. Tripathi, A., Backman, J., Elderfield, H. & Ferretti, P. *Nature* **436**, 341–346 (2005).
5. Coxall, H. K. et al. *Nature* **433**, 53–57 (2005).
6. Driscoll, N. W. & Haug, G. H. *Science* **282**, 436–438 (1998).
7. Delaney, M. L. & Boyle, E. A. *Paleoceanography* **3**, 137–156 (1988).
8. DeConto, R. M. & Pollard, D. *Nature* **421**, 245–249 (2003).
9. Zachos, J. C. & Kump, L. R. *Glob. Planet. Change* **47**, 51–66 (2005).
10. Kump, L. R., Kasting, J. F. & Crane, R. G. *The Earth System* 2nd edn (Prentice-Hall, Upper Saddle River, NJ, 2004).

BEHAVIOURAL GENETICS

Sex in fruitflies is *fruitless*

Charalambos P. Kyriacou

The courtship rituals of fruitflies are disrupted by mutations in the *fruitless* gene. A close look at the gene's products — some of which are sex-specific — hints at the neural basis of the flies' behaviour.

Richard Feynman is reported to have said, "Science is a lot like sex. Sometimes something useful comes of it, but that's not the reason we're doing it." In three papers, two published in *Cell*^{1,2}, and one in this issue³ (page 395), science and sex have come together to provide us with something useful — an extraordinary glimpse into how the male and female nervous systems function to generate sexual behaviour in fruitflies (*Drosophila*).

Unlike many British males on a Friday night, *Drosophila* males do not simply jump on the first female they see. Courtship behaviour in *D. melanogaster* is a stereotyped and instinctive sequence of behaviours performed by the male, involving visual, olfactory, gustatory, tactile, acoustic and mechanosensory stimuli being exchanged between the sexes (Fig. 1). The female's role is considerably less dramatic than the male's: she simply runs away, gives the odd kick, then mates (or not)⁴.

Normal mature males seldom court other males, but male *fruitless* mutants are bisexual, courting not only females but also other males⁵. In exclusive male company, these

mutants can form bizarre courtship chains, where several males, each chasing and courting the one in front, generate frenzied revolving circles.

The gene mutated in the *fruitless* flies (termed *fru*) was molecularly cloned in 1996, and the putative protein that it encodes was identified as a transcription factor^{6,7}, a regulatory molecule that controls gene expression. A large number of different messenger RNAs can be generated from the *fru* gene, some of which are sex-specific. In particular, an mRNA produced only in males is translated into a protein called Fru^M (for male-specific Fruitless)^{6,7}. This sex-specific production of the *fru* mRNAs is determined by the canonical sex-determination system, the most relevant component of which is the *transformer* gene (or *tra*).

Briefly, the encoded Tra protein binds to very short sequences (13 nucleotides) on the immature *fru* mRNA, to sex-specifically regulate which portions will be 'spliced' into the final transcript^{6,7}. (Indeed, Ryner *et al.*⁶ cloned *fru* by looking for genes that contained

Tra-binding sequences.) Similarly, Tra protein binds to the *doublesex* (*dsx*) gene and splices it in male- and female-specific modes (Dsx^M and Dsx^F, respectively)⁸. The Dsx^M and Dsx^F transcription factors mainly determine sexual morphologies⁸, but the sexual identity of the nervous system is shaped by *fru*.

By forcing males to express the female-specific *fruF* transcript, Demir and Dickson¹ produced males that showed the characteristics of the worst-affected *fru* mutants. These males were sterile, they barely courted females and they were more interested in courting males, forming courtship chains. By contrast, females jammed into *fruM* mode mated poorly, produced very few eggs, but — astonishingly — courted other females (Fig. 2), even to the point of forming chains. And an identity crisis of similar epic proportions was observed in females that were 'masculinized' using a different *fru*-related genetic trick³. Finally, by feminizing specific abdominal glands in males to produce female pheromones, and placing the altered males with *fruM* females, the sex roles were reversed, so that the females courted the males¹.

In another nifty piece of genetic engineering, both teams^{2,3} generated flies in which they could, among other things, mark the parts of the nervous system (just 2%) that show sex-specific expression of Fru. Further genetic manipulations showed that high levels of male–male courtship result when the communication between these neurons is shut down, or when *fruM* expression in these neurons in males is inhibited^{2,3}. Both studies found that the central nervous system of males and females looked very similar in terms of sex-specific *fru* expression, with few differences between the sexes in the numbers, positions or wiring of cells expressing Fru.

The *fru* products were found in almost all sensory organs that have been implicated in courtship^{2,3}. Olfactory sensory neurons showed some evidence for sexual dimorphisms. Those receptors that respond to pheromones project to certain other brain regions that are larger in males than females, reflecting the fact that sex pheromones have a greater functional significance in male *Drosophila*². By reversibly shutting down the *fru*-expressing olfactory receptors, both in males and in masculinized females in the



Figure 1 | The courtship sequence of *D. melanogaster* males. From left to right, the male orients towards the female, extends a wing and vibrates it, serenading the female with a species-specific love-song. He then licks the female's genitalia, attempts to copulate, and (maybe) copulates. (Drawings by B. Burnet.)

sex-reversal paradigm outlined above, courtship behaviour declined significantly, implying that these receptors are central to sexual interactions². However, by decreasing *Fru^M* in males just in these neurons, homosexual courtship increased, so normally these olfactory receptors must inhibit male–male interactions³.

So, a single *fru*-encoded genetic switch seems to be sufficient to shift the functioning of the nervous system from male to female mode, irrespective of the morphological sex of the animal. The general absence of large-scale sexual dimorphisms in *fru*-expressing neurons implies that it is the molecules regulated by *fru* that make the difference. Future work will undoubtedly be aimed at finding these molecules, as well as identifying the subset of key neurons that are sufficient to generate male courtship elements. Indeed, Vilella *et al.*⁹ have identified neurons downstream of ones expressing *fru* that are implicated in the control of the male's courtship song. Finally, an intriguing and mostly forgotten paper was published 30 years ago¹⁰ about 'lesbian' *Drosophila* females that courted



Figure 2 | Courtship remodelled. A *fru^M* female extends a wing as if 'singing' towards a normal female (reproduced with permission from ref. 1).

other females — apparently because of a genetic factor(s) on chromosome 2 (*fru* is on chromosome 3). Might this long-lost strain have carried a mutation in one of the *fru* target genes?

The work discussed here may well find itself

the focus of attention for those interested in the debate (scientific and political) on the genetic versus environmental bases of human sexuality. Perhaps we should remind ourselves that normal fly sexual preferences, unlike human sexual behaviour, cannot be modulated to any significant extent by altering experience¹¹.

Charalambos P. Kyriacou is in the Department of Genetics, University of Leicester, Leicester LE1 7RH, UK.

e-mail: cpk@leicester.ac.uk

1. Demir, E. & Dickson, B. J. *Cell* **125**, 785–794 (2005).
2. Stockinger, P., Kvitsiani, D., Rotkopf, S., Tirian, L. & Dickson, B. J. *Cell* **125**, 795–807 (2005).
3. Manoli, D. S. *et al.* *Nature* **436**, 395–400 (2005).
4. Greenspan, R. J. & Ferveur, J. -F. *Annu. Rev. Genet.* **34**, 205–232 (2000).
5. Vilella, A. *et al.* *Genetics* **147**, 1107–1130 (1997).
6. Ryner, L. C. *et al.* *Cell* **87**, 1079–1089 (1996).
7. Ito, H. *et al.* *Proc. Natl Acad. Sci. USA* **93**, 9687–9692 (1996).
8. Cline, T. W. & Meyer, B. J. *Annu. Rev. Genet.* **30**, 637–702 (1996).
9. Vilella, A., Ferri, S. L., Krystal, J. D. & Hall, J. C. *Proc. Natl Acad. Sci. USA* (in the press).
10. Cook, R. *Nature* **254**, 241–242 (1975).
11. Siegel, R. W., Hall, J. C., Gailey, D. A. & Kyriacou, C. P. *Behav. Genet.* **14**, 383–410 (1984).

ASTEROIDS

Shaken on impact

Erik Asphaug

A single recent impact may have modified the craters on the asteroid Eros into the pattern we see today. This finding has implications for how we view the structure of asteroids — and for addressing any hazards they present.

Asteroids seem to get stranger with every passing year. Thomas and Robinson's finding (page 366 of this issue)¹ — that impact-induced vibrations of an asteroid may be the dominant mechanism reshaping its surface — shakes things up still further. In the case of the well-studied asteroid Eros, the authors link this resurfacing mechanism to the recent impact of a meteoroid that left a particularly large crater. They thereby make the first detailed mechanical connection between surface observations and an asteroid's global geology. The authors conclude that Eros, a rocky asteroid 33 by 13 by 13 kilometres in size, has a relatively homogeneous interior that transmits seismic shocks efficiently and is mantled by a hundred metres or more of regolith. (Regolith is the loose soil-like material familiar from pictures of the surface of the Moon.) This might not come as a surprise, given Eros's appearance², but for the first time, the authors provide convincing evidence that makes this conclusion more than just reasonable conjecture.

Thomas and Robinson's discovery marks another stage in the journey asteroids have taken from insignificance, through notoriety, into the mainstream of scientific interest. The turning point came in the 1980s, when an asteroid was found to be responsible for the

greatest calamity to befall Earth's biosphere since the Permian era — an impact 65 million years ago in present-day Mexico that is postulated, among other things, to have wiped out the dinosaurs. That got people's attention. But the geological subtleties of asteroids remained largely unappreciated for a further ten years. This situation began to change with the first detailed ground-based radar observations³, and the Galileo mission's fly-by of the asteroids Ida and Gaspra⁴. Now, a new generation of scientists is appreciating asteroids as geological entities^{2,5,6}.

If Thomas and Robinson's hypothesis of seismic shaking¹ is correct, then the cratering history of any asteroid is complex. Impacts of small meteoroids make the surface heavily cratered, giving it an 'old' look, whereas impacts of larger meteoroids — by causing the surface to vibrate — erase smaller craters, making the asteroid appear 'young'. This asteroidal Botox calls into question the habit of dating asteroid surfaces through their cratering record: although the passage of time is indeed recorded here, so too is internal structure. A young asteroid of the type that resembles a rubble pile, for instance, is more capable of damping vibrations, and might retain more craters — and so appear older — than an

ancient, 'competent' asteroid that has a more monolithic interior and thus transmits seismic energy more effectively. But Thomas and Robinson's work also opens up a new way of looking at asteroids generally. It shows how we might gauge interior structure from surface observations: craters and other landforms, and their degradation, could be used as proxies for seismic data.

The idea of seismic processes resurfacing asteroids is not itself new. The formation of the large crater Stickney on Phobos (Fig. 1, overleaf), a martian moon about the size of Eros and perhaps a captured asteroid, was modelled⁷ 12 years ago using a computational tool called a hydrocode to simulate the effect of the high-velocity impact. The simulation showed that seismic resurfacing could erase craters smaller than about 100 metres in diameter, and significantly degrade larger craters. The same method was later used to show⁸ that the jolting of the asteroid Gaspra by large impacts could lead to the unusual distribution of its crater sizes. In an argument analogous to that used by Thomas and Robinson for Eros, the asteroid Ida was suggested⁹ to have a relatively monolithic deep interior, given evidence that stress energy was transmitted from a large impact structure at one end to



50 YEARS AGO

'Automation' was defined by R. K. Geiser at a conference on automation and industrial development at Syracuse, New York, in May 1954, as "the accomplishment of a job by an integrated mechanism with a minimum of assistance of any kind". Full automation, in this sense, does not yet exist, though it may be achieved in a suitable industry before the end of this century. Any technical development that enables a machine or instrument to dispense with labour is a step in this direction; but whereas mechanization has largely displaced physical effort, the new developments are making automatic some of the work that was formerly done by human brains... the implications in respect of leisure and its use [should not] be overlooked... increased productivity is likely to be accompanied by an increase in leisure, and this can only be detrimental to a community that is unprepared or unable to make profitable use of that leisure.

From *Nature* 23 July 1955.

100 YEARS AGO

Commander Peary sailed on Sunday last to make a further attempt to reach the North Pole. Before leaving, he communicated various particulars respecting his expedition to Reuter's Agency. His plan is based upon the Smith Sound, or "American" route to the Pole, and his object is to force his ship to a base within 500 miles of the Pole itself, and then to sledge across the Polar pack. The Arctic ship *Roosevelt*, which has been specially built for this expedition, has been constructed so as to withstand the heavy ice pressure and is so shaped that the pressure of the ice pack will have the effect of raising the vessel out of the water. The ship will carry a wireless telegraphic outfit, which, with one or two relay stations in Greenland, will keep her in communication with the permanent telegraph station at Chateau Bay, Labrador, and thence by existing lines with New York.

From *Nature* 20 July 1905.



Figure 1 | Captured asteroid. The pitted surface of the martian moon Phobos, about the size of Eros — the subject of Thomas and Robinson's study¹ — and thought to be an asteroid caught by Mars's gravitational pull. The giant crater Stickney, 10 kilometres in diameter, is clearly visible at the top of the picture. Modelling of the impact that formed this crater⁷ led to the idea that seismic processes can resurface asteroids — a process now examined in detail¹¹ that may help us to determine the mechanical structure of asteroids.

form a series of impact-induced fracture grooves at the other. And recently, the most detailed seismological model for asteroids so far was developed by James Richardson and colleagues¹⁰ at the University of Arizona to explain the lack of small craters on Eros.

So the stage was already set for Thomas and Robinson. They seek¹ to make specific correlations and so explain why certain areas of Eros are almost devoid of small craters, whereas other, often adjacent, areas are heavily cratered. They argue that these discontinuities in crater density (Fig. 2 of ref. 1, page 367) cannot be interpreted as patterns caused by debris thrown out during the formation of other, larger craters; nor are they easily explained by any continuum degradation process. They apply a model for seismic shaking in which the vibration of Eros's surface drops off as a simple function of distance from a single specific impact site. This admittedly naive approach provides a surprisingly good fit to the observed density of small craters, assuming that Eros is able to transmit stresses efficiently throughout its interior, and is covered in loose material about 100 metres in depth. Such assumptions are in agreement with earlier geological interpretations for Eros^{2,6} and also for Ida⁹, an asteroid quite similar in outward appearance.

These first steps in the 'passive seismology' of asteroids are particularly encouraging as we move rapidly towards a new era of space exploration. This era has in no small part been compelled by the discovery of many small bodies, asteroids and comets, collectively known as near-Earth objects or NEOs, that pass alarmingly close to us. Over the years this list is likely to be narrowed to a few dozen

seriously hazardous asteroids. Still, it will certainly be useful to say more about these objects — which could potentially become as notorious as Vesuvius or Popocatepetl — than their equivalent destructive power in millions of megatonnes of explosive. A little knowledge could go a long way to ensure that disaster movies remain in the realm of science fiction.

There is a lot more to be done before surface observations can be used to provide definitive knowledge about asteroid interiors: direct data on interior structure, derived from radar or seismological observations, are almost certainly required to validate these ideas. Even now, as we sift through the debris from the encounter of the space-probe Deep Impact with comet Tempel 1 (ref. 11), one cannot help but think that this excellent experiment would be perfectly complemented by a couple of seismological geophones anchored to the surface of the comet. Those who recall the heyday of lunar spaceflight remember that we began learning about the Moon by crashing things into it, too.

Erik Asphaug is in the Department of Earth Sciences, University of California, 1156 High Street, Santa Cruz, California 95064, USA. e-mail: asphaug@pmc.ucsc.edu

1. Thomas, P. C. & Robinson, M. S. *Nature* **436**, 366–369 (2005).
2. Robinson, M. S., Thomas, P. C., Veverka, J., Murchie, S. L. & Wilcox, B. B. *Meteorit. Planet. Sci.* **37**, 1651–1684 (2002).
3. Ostro, S. J. *Rev. Mod. Phys.* **65**, 1235–1279 (1993).
4. Belton, M. J. S. et al. *Science* **265**, 1543–1547 (1994).
5. Sullivan, R. et al. *Icarus* **120**, 119–139 (1996).
6. Prockter, L. et al. *Icarus* **155**, 75–93 (2002).
7. Asphaug, E. & Melosh, H. J. *Icarus* **101**, 144–164 (1993).
8. Greenberg, R. et al. *Icarus* **107**, 84–97 (1994).
9. Asphaug, E. et al. *Icarus* **120**, 158–184 (1996).
10. Richardson, J. E., Melosh, H. J. & Greenberg, R. *Science* **306**, 1526–1529 (2004).
11. Peplow, M. *Nature* **436**, 158–159 (2005).

METABOLISM

A is for adipokine

Deborah M. Muoio and Christopher B. Newgard

Adipokines are hormones that signal changes in fatty-tissue mass and energy status so as to control fuel usage. A fat-derived adipokine that binds to vitamin A provides a new link between obesity and insulin resistance.

The worldwide epidemic of obesity has been accompanied by a surge in the incidence of diabetes¹. Normally, control of blood glucose levels depends on the efficient action of insulin, which stimulates uptake of glucose from the blood and slows its output from the liver. In both obesity and diabetes, target tissues such as muscle and liver fail to adjust glucose metabolism appropriately in response to insulin. The onset of this 'insulin-resistant' condition is intimately associated with weight gain¹, suggesting that increased fatty adipose tissue generates a signal (or signals) that interferes with the action of insulin. Consistent with this notion, in this issue Yang *et al.* (page 356)² report that a factor derived from fat cells, called retinol binding protein-4 (RBP4), can impair insulin sensitivity

throughout the body. RBP4 joins a growing list of fat-derived peptides that modulate glucose homeostasis.

The significance of adipose tissue as an endocrine organ first surfaced in 1995 with the ground-breaking discovery of leptin³. This fat-derived hormone controls body weight by regulating both feeding behaviour and energy expenditure. Ensuing research uncovered a whole family of adipose-derived 'adipokines' (for example, adiponectin, TNF- α , resistin) that signal changes in the mass of adipose tissue and energy status to other organs that control fuel usage⁴. From a clinical viewpoint, each of these secreted peptides represents a possible drug target with the potential to uncouple insulin resistance from obesity.

Yang and colleagues' findings² may provide the solution to a long-standing paradox in diabetes research. The expression of GLUT4, an insulin-regulated glucose transporter, is greatly reduced in the fat cells (adipocytes) but not in the muscle cells of rodents and humans that are obese and have insulin resistance⁵. This is surprising given the predominant role of muscle in the disposal of glucose. The first clues to solving this puzzle emerged from studies in which the expression of GLUT4 was either ablated or increased specifically in adipose tissue^{6,7}. Mice lacking GLUT4 in their adipose tissue are prone to diabetes⁸, whereas those with overexpression of GLUT4 exhibit increased efficiency of glucose clearance⁶. These changes in whole-body insulin action occur through alterations in the sensitivity of muscle and liver cells to insulin, thereby implicating an 'adipocrine' substance that allows fat to communicate with peripheral tissues. However, a survey of the known adipose-derived factors, including leptin, free fatty acids and TNF- α , failed to reveal a candidate that responded to the GLUT4 manipulations in a convincing manner.

Now Yang *et al.* have used DNA microarrays to search for other adipokines. They identified RBP4 as a secreted protein that is regulated

PARASITOLOGY

Triple genome triumph

There is welcome news for scientists working on sleeping sickness, Chagas' disease and visceral leishmaniasis: the genomes of the three trypanosome parasites responsible for these devastating illnesses have now been cracked. The sequences from *Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania major* were published in last week's *Science* by an array of international research teams (*Science* **309**, 416–422, 409–415, 436–442; 2005).

In the terminology of global public health, these diseases don't even fall into the category of 'neglected diseases' such as malaria and tuberculosis. Rather, they are classed as 'most neglected diseases' — which nonetheless kill millions. But those affected have little means of paying for treatment, making drug development unprofitable. Consequently, there are no vaccines, and medicines are few, expensive and usually toxic.

Treatment of sleeping sickness, for example, still relies on melarsoprol, a 50-year-old drug that is ineffective in a third of patients and kills 5% of those who take it. The high rate of

fatal reactions is accepted because the disease is otherwise lethal. New therapies are clearly needed, and the availability of the parasite genomes is a step towards finding drug targets and vaccine candidates.

The three parasites share around 6,200 'core' genes, so the proteins these encode might provide targets for drugs that are effective against all three. The parasites make a large and diverse set of kinase and phosphatase enzymes. This means that there could well be regulatory and other processes used by the organisms that could be vulnerable to disruption by drugs.

Many species-specific genes were also identified in the genome sequences, providing potential species- and stage-specific targets. Although the three parasites share many subcellular structures, such as kinetoplasts and glycosomes, the organisms are very different. They are spread by different insects, attack different tissues and cause different pathologies. The specimens of *L. major* pictured are in the form that is transmitted to humans by sand flies.



Each parasite also has its own mechanism for evading the human immune system: *T. brucei* does not enter its victim's cells, and evades the immune system by constantly changing its main surface proteins; *T. cruzi* holes up inside cells, but uses a similar strategy to hide from the immune system; and *L. major* infects certain immune cells and interferes with their function.

Producing effective treatments

against these parasites will be a lengthy process, but initial research is already under way by not-for-profit drugs groups such as the Institute for OneWorld Health (www.oneworldhealth.org) and the Drugs for Neglected Diseases Initiative (www.dndi.org). The genome sequences will provide such initiatives with a wealth of data and leads.

Declan Butler

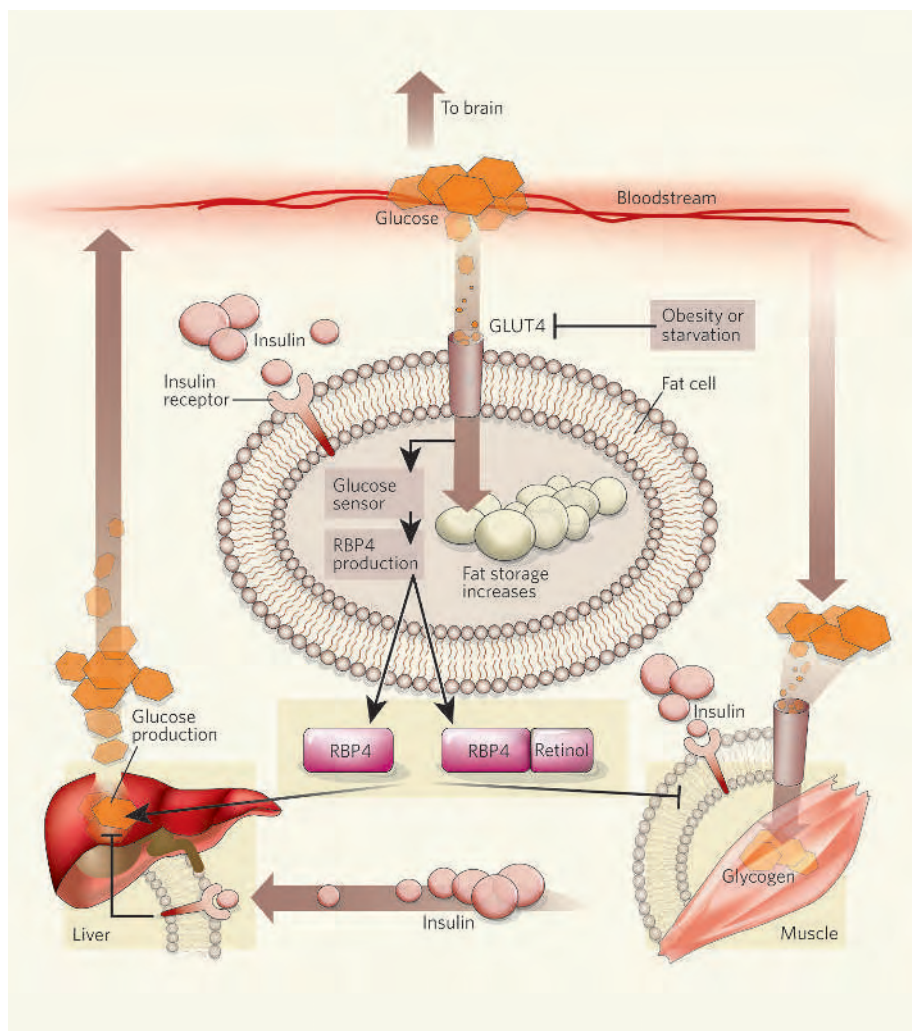


Figure 1 | Retinol binding protein-4 (RBP4) in glucose metabolism. In normal individuals, binding of insulin to its receptor on the cell membrane stimulates glucose uptake into muscle and fat cells through the GLUT4 transporter. It also inhibits glucose production in liver, thereby maintaining normal glucose levels in the blood. In adipose tissue, glucose provides fuel for the synthesis of fat stores, which serve as the body's main energy reservoir. Yang *et al.*² found that the decrease in GLUT4 expression that occurs in the fatty tissue of obese animals is accompanied by increased expression and secretion of the fat-derived factor RBP4. This factor, possibly working in concert with retinol (vitamin A), impairs insulin signalling in muscle, inhibiting glucose uptake, and interferes with insulin-mediated suppression of glucose production in the liver, causing blood glucose levels to rise.

reciprocally in adipose tissue of mice over-expressing GLUT4 and those lacking GLUT4. The authors provide comprehensive support for RBP4 as the elusive link between GLUT4 suppression in adipose tissue and insulin resistance, and as a central mediator of insulin action generally (Fig. 1). Circulating RBP4 levels are raised in five independent mouse models of obesity and insulin resistance, as well as in obese humans. In mice lacking GLUT4, a drug used to treat diabetes, rosiglitazone, lowers circulating RBP4 levels and normalizes insulin sensitivity. Increasing the circulating levels of RBP4 results in glucose intolerance, and conversely, deleting the RBP4 gene in mice increases insulin sensitivity. Finally, Yang *et al.* show that treatment of mice with the synthetic retinoid fenretinide, which increases the excretion of RBP4, lowering its levels in the blood, ameliorates insulin

resistance caused by high-fat feeding.

The mechanisms by which RBP4 affect insulin action are partially elucidated. In muscle, RBP4 decreases the activity of the enzyme PI-3 kinase and the phosphorylation of insulin receptor substrate-1, both effects being clear markers of impaired insulin action. Increasing RBP4 does not alter PI-3 kinase activity in liver, yet glucose production in the liver is clearly increased, in concert with increased expression of a key enzyme in the glucose production pathway, phosphoenolpyruvate carboxykinase (PEPCK).

As its name suggests, RBP4 was previously known as a transporter for retinol (vitamin A)⁹. It is unclear whether the link between RBP4 and insulin action involves changes in retinol metabolism or delivery. Of interest in this context is the fact that PEPCK expression is stimulated by retinoids, an effect that could

be mediated by enhanced delivery of the retinol ligand by RBP4. However, Yang *et al.*² also show that RBP4 stimulates PEPCK expression and glucose production in cultured rat cells, which could imply that the peptide has a direct effect, although no high-affinity receptor for RBP4 has been identified. In addition, retinol serves as a precursor for the synthesis of ligands of the RAR and RXR nuclear hormone receptors¹⁰. RXR is a partner for a family of receptors (peroxisome-proliferator-activated receptors) that regulate transcription of genes involved in fatty-acid metabolism¹¹. RBP4 might therefore be related to diabetes through dysregulation of intramuscular and/or hepatic fatty-acid metabolism, a well-recognized component of insulin resistance¹².

Finally, did the adipocyte GLUT4–RBP4 system evolve as a response to obesity and overeating, or for other purposes? Interestingly, food deprivation (for example, overnight fasting) promotes insulin resistance, and dramatically reduces GLUT4 expression in adipose tissue¹³. Whether RBP4 levels rise in response to this mode of GLUT4 regulation is not known. But if they do, this might indicate that the GLUT4–RBP4 system evolved as a mechanism for restricting glucose uptake by peripheral tissues under famine conditions, thereby sparing glucose for the brain, which depends on the sugar as its primary energy source. An early consequence of obesity, in contrast, may be development of insulin resistance in the adipocyte. This would result in the same fall in GLUT4 expression that occurs during fasting, causing the adipocytes in essence to mistake obesity for starvation. Clearly, the study by Yang *et al.*² moves the adipocyte and its secreted factors closer to the epicentre of the diabetes and obesity epidemic.

Deborah M. Muoio and Christopher B. Newgard are in the Sarah W. Stedman Nutrition and Metabolism Center, and the Departments of Pharmacology and Cancer Biology and of Medicine, Duke University Medical Center, Durham, North Carolina 27710, USA.

e-mail: newga002@mc.duke.edu

- Engelgau, M. M. *et al.* *Ann. Intern. Med.* **140**, 945–950 (2004).
- Yang, Q. *et al.* *Nature* **436**, 356–362 (2005).
- Halaas, J. L. *et al.* *Science* **269**, 543–546 (1995).
- Mora, S. & Pessin, J. E. *Diabetes Metab. Res. Rev.* **18**, 345–356 (2002).
- Kahn, B. B. *J. Nutr.* **124**, 1289S–1295S (1994).
- Tozzo, E., Shepherd, P. R., Gnudi, L. & Kahn, B. B. *Am. J. Physiol. Endocrinol. Metab.* **268**, E956–E964 (1995).
- Gnudi, L., Shepherd, P. R. & Kahn, B. B. *Proc. Nutr. Soc.* **55**, 191–199 (1996).
- Abel, E. D. *et al.* *Nature* **409**, 729–733 (2001).
- Quadro, L., Hamberger, L., Colantuoni, V., Gottesman, M. E. & Blaner, W. S. *Mol. Aspects Med.* **24**, 421–430 (2003).
- Marill, J., Idres, N., Capron, C. C., Nguyen, E. & Chabot, G. G. *Curr. Drug Metab.* **4**, 1–10 (2003).
- Ferre, P. *Diabetes* **53** (Suppl. 1), S43–S50 (2004).
- McGarry, J. D. *Diabetes* **51**, 7–18 (2002).
- Sivitz, W. I., Desautel, S. L., Kayano, T., Bell, G. I. & Pessin, J. E. *Nature* **340**, 72–74 (1989).

BRIEF COMMUNICATIONS

Jewish inspiration of Christian catacombs

A Jewish cemetery in ancient Rome harbours a secret that bears on the history of early Christianity.

The famous catacombs of ancient Rome are huge underground cemeteries, of which two Jewish catacomb complexes of uncertain age and 60 early-Christian catacombs have survived^{1–3}. Here we use radiocarbon dating to determine the age of wood originating from one of the Jewish catacombs and find that it pre-dates its Christian counterparts by at least 100 years. These results indicate that burial in Roman catacombs may not have begun as a strictly Christian practice, as is commonly believed^{1,3,4}, but rather that its origin may lie in Jewish funerary customs.

The Jewish and Christian catacombs of Rome are all thought to date from the same general period — namely, from the early third century through to the early fifth century AD (refs 2,5) — but more precise dating has been difficult^{6,7}. We therefore collected organic material for radiocarbon dating that had been incorporated during the construction of grave recesses (loculi) in the Jewish Villa Torlonia catacomb^{8,9} (Fig. 1, and see supplementary information).

Each loculus was sealed by a small wall of rubble and bricks that was covered in a smooth layer of lime. Pieces of charcoal from the limekiln that had been embedded in the lime were collected for dating. As young trimmings would have been the preferred wood for generating the high temperature (around 900 °C) necessary to effect the conversion of limestone to lime, we consider that contamination from mature or already dead wood is unlikely.

Fifteen samples from the five interconnected regions (A–E in Fig. 1) that make up the catacomb were subjected to radiocarbon analysis by atomic-mass spectroscopy¹⁰; absolute ages were determined for 2σ ranges by using the computer code Calib4 (ref. 11). Results from five damaged samples had to be discarded, but preliminary analysis of the remainder looked promising¹².

Our analysis reveals a complete construction history of the upper and lower catacombs. The sequence of 2σ calendar ages ranges from 50 BC (from a sample at the entrance area to the lower catacomb) to AD 400 (sample from region A in the upper catacomb) (see supplementary information). Various charcoal samples were systematically identified as being derived from taxa known to be growing in the area during the time of the catacomb's construction¹³ (see supplementary information).

Our results are consistent with the chronological layout of the catacomb (Fig. 1). The



Figure 1 | Layout of upper and lower Jewish catacombs in Villa Torlonia in Rome. For radiocarbon dates of samples taken from different construction sites, see supplementary information. Numbers refer to the sample origin; letter-number combinations refer to gallery number. Scale bar, 10 m.

oldest sample (number 1) derives from the entrance, the earliest construction point. Samples 3–6 from the D region fit the topographical development of the catacomb; the age of a painted arcosolium (arched grave) in the upper catacomb (sample number 7) is consistent with dating proposed on the basis of its style (AD 320–350; ref. 5). The only sample that does not fit our chronology is number 10, which originates from a gallery that was dug into a pre-existing network of underground water tunnels.

This evidence indicates that the Villa Torlonia catacomb came into use in the second century AD, a century before the building of the earliest Christian catacombs started. Given that Roman Christianity evolved from Judaism, and Jews and Christians continued to interact until well into Late Antiquity, it is possible that Christian funerary practices were influenced by Jewish ones. This could explain the similarity between the oldest of the early Christian underground cemeteries and the Jewish Villa Torlonia catacomb, particularly considering that Callixtus, the deacon in charge of developing the Christian catacombs, came from the Jewish quarter. However, confirmation awaits radiocarbon dating of the Christian catacombs.

Leonard V. Rutgers*, Klaas van der Borg†, Arie F. M. de Jong†, Imogen Poole‡

*Faculty of Theology, Utrecht University,

3508 TC Utrecht, The Netherlands

e-mail: lrutgers@theo.uu.nl

†AMS Facility Utrecht, Subatomic Physics

Department, Utrecht University,

3508 TA Utrecht, The Netherlands

‡National Herbarium Nederland, Utrecht University Branch, 3585 CS Utrecht, The Netherlands

1. Flocchi Nicolai, V. & Bisconti, F. & Mazzoleni, D. *The Christian Catacombs of Rome: History, Decorations, Inscriptions* (Schnell and Steiner, Regensburg, 1999).
2. Rutgers, L. V. *Subterranean Rome* (Peeters, Leuven, 2000).
3. Flocchi Nicolai, V. *Strutture Funerarie ed Edifici di Culto Paleocristiani di Roma dal IV al VI Secolo* (Pontificia Commissione di Archeologia Sacra, Vatican City, 2001).
4. Pergola, P. *Le Catacombe Romane. Storia e Topografia* (Nuova Italia Editrice, 1997).
5. Rutgers, L. V. *The Hidden Heritage of Diaspora Judaism* (Leuven, Peeters, 1998).
6. Guyon, J. *Boreas* **17**, 89–103 (1994).
7. Deckers, J. G. *Studi di Antichità Cristiana* **48**, 217–238 (1992).
8. Fasola, U. M. *Rivista di Archeologia Cristiana* **52**, 7–62 (1976).
9. Barbera M. & Magnani Cianetti, M. in *I Beni Culturali Ebraici in Italia. Situazione Attuale, Problemi, Prospettive e Progetti per il Futuro* (ed. Perani, M.) 55–70 (Longo Editore, Ravenna, 2003).
10. van der Borg, K. et al. *Nucl. Instr. Meth.* **B123**, 97–101 (1997).
11. Stuiver, M. & Reimer, P. J. *Radiocarbon* **35**, 215–230 (1993).
12. Rutgers, L. V., van der Borg, K. & de Jong, A. F. M. *Radiocarbon* **44**, 541–547 (2002).
13. Fenaroli, L. *Gli Alberi d'Italia* (Aldo Martello Editore, Milan, 1967).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/436339a

BRIEF COMMUNICATIONS ARISING online
www.nature.com/bca see Nature contents.

PALAEOCLIMATOLOGY

Formation of Precambrian sediment ripples

Arising from: P. A. Allen & P. F. Hoffman *Nature* **433**, 123–127 (2005)

Quantitative estimation of environmental properties using sedimentary structures preserved in rocks is complicated by the fact that some relationships between the fluid flow, sediment transport and bed topography are not unique. Allen and Hoffman¹ propose that large, wave-generated sand ripples (orbital ripples) in Precambrian rocks were generated by sustained, extreme winds driven by rapid climate change after termination of the Marinoan glaciation. We show here that these features could equally well have formed under normal storm conditions in tens of metres of water. We therefore contend that the ripples do not provide direct evidence for a climatic transit after the break-up of a snowball-Earth's global ice cover.

Allen and Hoffman conclude that the observed ripples developed in deep water (depth h , 200–400 m), by waves of unusually large period (T , 21–30 s) and amplitude (H , 7.5–12 m). They suggest that a discrete cyclone or hurricane is likely to be too short-lived an event to produce the observed sedimentary structures and that present-day orbital ripples seldom have wavelengths (λ) exceeding 1 m, both of which we contest.

A bathymetric survey² of the continental shelf off North Carolina in the United States found ripples with wavelengths of up to 4 m and a median grain diameter (D) of 0.1–5 mm covering the shelf at h values of 20–40 m. Field observations link formation of these ripples to specific hurricanes and tropical storms. Measured values of T and H for the water waves developed during these events commonly exceeded 60 s and 3 m, respectively. Detection of this bed topography seems to be limited by instrument resolution, rather than by a paucity of these features on the sea floor².

Critical shear stress (ψ_c) for the initial motion of a particle of size D provides a minimum bed-stress condition for ripple formation^{1–3}. An upper limit for bed stress associated with steep orbital ripples is $3\psi_c$ (ref. 2). This narrow range in bed shear stress plus the mean value for λ constrain the associated near-bed flow field^{2,3} (Fig. 1). With these parameters, T is the only surface-wave property that can be estimated from sedimentary deposits^{1,3}. Airy wave theory relates wavelength (L), H and h to near-bed flow conditions^{1,3}; however, an infinite combination of these variables can produce the same near-bed conditions (Fig. 1). Allen and Hoffman only consider transport conditions at $\psi = \psi_c$, which yields a maximum estimate for T . For $\psi = 3\psi_c$, T is reduced by a factor of $\sqrt{1/3}$ (Fig. 1). Flanks of some preserved ripples exceed the angle of repose, indicating deformation and

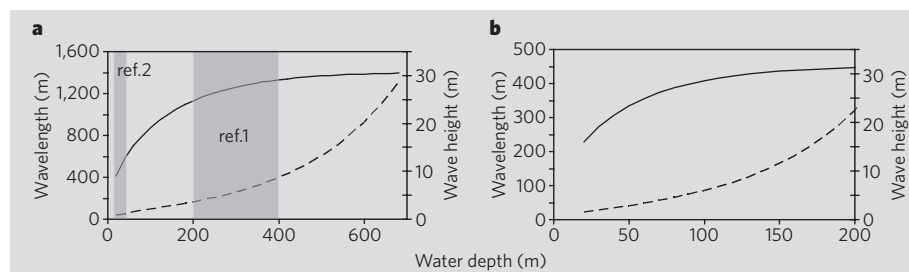


Figure 1 Height (dashed lines) and length (full lines) of possible ripple-forming surface waves calculated for a range of water depths. **a**, **b**, Calculations using Airy theory^{2,3} with **a**, $T = 30$ s, $\psi = \psi_c$, and **b**, $T = 17$ s, $\psi = 3\psi_c$, where T is the wave period and $\psi(\psi_c)$ is the (critical) shear stress. Ranges of wave conditions associated with reported water depths from refs 1 and 2 are indicated by grey boxes. For the figure, we used wavelength $\lambda = 3.5$ m and ψ_c estimated from grain size $D = 0.12$ mm (for methods, see ref. 2). Using entire ranges of D (0.12–0.5 mm) and λ (1.5–5.4 m) reported by Allen and Hoffman, and $\psi_c \leq \psi \leq 3\psi_c$ (ref. 2), yields T values of 8–41 s. We calculated the aggradation rate as $r = m/T$, where $m = 0.40$ cm is the mean cross-bed thickness in Fig. 2c of ref. 1. When T is between 8 and 41 s, r is 0.58–2.99 cm min⁻¹.

making the measured steepness values inexact.

More important, Allen and Hoffman assume, without justification, that wind of unlimited fetch and duration generated the long-period surface waves producing the bed-forms. Their estimates for H are based on this model and these values, in turn, are used to calculate h . Their environmental reconstruction represents a possible, but non-unique, inversion of the geological data. Modern storm-generated waves of similar period produce orbital ripples of the same morphology and grain size² as the Marinoan examples, but under conditions of much smaller h , H and L . An independent constraint on any one of these three variables is necessary for closure. The most reasonable procedure would be to estimate water depth based on the physiographical position of ripples found within the ancient basin.

Perhaps the most remarkable aspect of the reported stratigraphy¹ is the continuous vertical climb of the ripples. A rate of deposition associated with this climb is tightly constrained by T , and is calculated to be about 1 cm min⁻¹. This high rate seems to rule out spontaneously precipitating carbonate⁴ as the sediment source for the ripples. At this rate, the entire sequence

shown in Fig. 3 of ref. 1 could have been deposited in less than 3 h. A small number of short-duration events do not place any constraint on associated climate conditions.

Our results (Fig. 1) show that the preserved orbital ripples¹ could have formed under rather mundane environmental conditions², and therefore do not provide evidence for extreme climate change. We wish to make clear that our analysis does not address larger issues of the snowball-Earth hypothesis⁴, but rather serves to show that small-scale observations must be carefully placed within a basin-scale context to produce a unique set of environmental conditions associated with the accumulation of the observed sedimentary deposits.

Douglas J. Jerolmack, David Mohrig

Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
e-mail: douglasj@mit.edu

1. Allen, P. A. & Hoffman, P. F. *Nature* **433**, 123–127 (2005).
2. Ardhuin, F. et al. *J. Geophys. Res.* **107**, 3143 (2002).
3. Komar, P. D. & Miller, M. C. *J. Sedim. Petrol.* **43**, 1101–1110 (1973).
4. Hoffman, P. F. et al. *Science* **281**, 1342–1346 (1998).

doi:10.1038/nature04025

PALAEOCLIMATOLOGY

Allen and Hoffman reply

Reply to: Jerolmack, D. J. & Mohrig, D. *Nature* doi:10.1038/nature04025 (2005)

Jerolmack and Mohrig¹ suggest that the wave-ripple structure we describe² from cap carbonates deposited in the aftermath of the Marinoan glaciation was created under storms or hurricanes similar to those experienced

today on certain oceanic coasts, citing a documentation of large wave ripples on the seabed off the coast of North Carolina. Side-scan sonar images indicate that such ripples have wavelengths of 0.4–3 m, although the ripple

dimensions at sites where samples were obtained for grain-size analysis range from 0.77 to 1.37 m, which is somewhat smaller than the Neoproterozoic examples shown in Table 1 of ref. 2 (1.5–4.5 m). In addition, the side-scan sonar equipment could only be deployed during fair-weather conditions after the passage of several hurricanes, which makes the precise hydraulic conditions responsible for the wave ripples uncertain. Nevertheless, Jerolmack and Mohrig raise an important issue regarding the shear stress (or orbital velocity) required to generate the wave ripples.

We argue that, at values of shear stress well above the threshold condition, vortex ripples lose their trochoidal profiles, flatten in steepness and become three-dimensional³. Only very steep wave ripples with trochoidal crests were used in our palaeohydraulic analysis, and other ripples of very large wavelength but lower steepness were excluded. A compilation of 648 self-consistent sets of laboratory and field data⁴ indicates that at a grain size in the range 0.12–0.5 mm, ripples of steepness of about 0.25 should be constrained within a very narrow field close to the threshold condition. At lower values of steepness, the existence field of vortex ripples expands to a broader range of orbital velocity. All the large metre-spacing wave ripples generated in purely oscillatory flow experiments at long periods in closed

ducts with 0.19–0.3-mm sand have markedly rounded crestal regions and a strong tendency towards three-dimensionality⁵. The discussion therefore centres on the existence field of steep (about 0.25), large-wavelength (more than 1.5 m), trochoidal wave ripples.

We agree that small-scale observations must be carefully placed in a wider context. This wider context is that the large wave ripples occur at the same stratigraphic position within the cap dolostone on five present-day continents. Either the winds and waves operated over a wide palaeogeographical belt, in which case a very short time period for wave-ripple formation is plausible, or they were generated under the spatially limited tracks of hurricanes, in which case we require much longer periods of time to integrate the hurricane activity over a large enough area to leave a widespread imprint at the same stratigraphic level. In addition, the azimuths of the wave-ripple crestlines vary little from bed to bed within the cap dolostone at any given location, supporting the idea of sustained zonal winds rather than superimposed hurricane tracks.

The time period for the vertical growth of the wave ripples is calculated by Jerolmack and Mohrig¹ on the assumption that each lamina represents deposition during one half-cycle of wave motion. As each lamina is of the order of 1–2 mm thick, we agree that the aggradational

ripples could have been deposited in a period of several hours, although this would require the entire climbing structure to be due to continuous aggradation. The deposition of 1.5 m of carbonate sediment in a period of several hours is itself testimony to extreme conditions. What cannot be proved on the basis of the palaeohydraulic analysis alone is whether the structures formed under intense hurricanes in relatively shallow water, or in deeper water under more sustained but unsteady zonal winds during climatic transit. Detailed examination of wave-generated structures in Marinoan cap dolostones worldwide will help to resolve this issue.

Philip Allen*, Paul Hoffman†

*Department of Earth Sciences, ETH Zürich, 8092 Zürich, Switzerland

e-mail: philip.allen@erdw.ethz.ch

†Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138-2902, USA

1. Jerolmack, D. J. & Mohrig, D. *Nature* doi:10.1038/nature04025 (2005).
2. Allen, P. A. & Hoffman, P. F. *Nature* **433**, 123–127 (2005).
3. Nielsen, P. J. *Geophys. Res.* **86**, 6467–6472 (1981).
4. Allen, J. R. L. *Dev. Sedimentol.* **30**, 444–448 (1984).
5. Southard, J. B., Lambie, J. M., Federico, D. C., Pile, H. T. & Wiedman, C. R. *J. Sedim. Petrol.* **60**, 1–17 (1990).

doi:10.1038/nature04026

Eocene bipolar glaciation associated with global carbon cycle changes

Aradhna Tripathi¹, Jan Backman², Henry Elderfield¹ & Patrizia Ferretti¹

The transition from the extreme global warmth of the early Eocene 'greenhouse' climate ~55 million years ago to the present glaciated state is one of the most prominent changes in Earth's climatic evolution. It is widely accepted that large ice sheets first appeared on Antarctica ~34 million years ago, coincident with decreasing atmospheric carbon dioxide concentrations and a deepening of the calcite compensation depth in the world's oceans, and that glaciation in the Northern Hemisphere began much later, between 10 and 6 million years ago. Here we present records of sediment and foraminiferal geochemistry covering the greenhouse-icehouse climate transition. We report evidence for synchronous deepening and subsequent oscillations in the calcite compensation depth in the tropical Pacific and South Atlantic oceans from ~42 million years ago, with a permanent deepening 34 million years ago. The most prominent variations in the calcite compensation depth coincide with changes in seawater oxygen isotope ratios of up to 1.5 per mil, suggesting a lowering of global sea level through significant storage of ice in both hemispheres by at least 100 to 125 metres. Variations in benthic carbon isotope ratios of up to ~1.4 per mil occurred at the same time, indicating large changes in carbon cycling. We suggest that the greenhouse-icehouse transition was closely coupled to the evolution of atmospheric carbon dioxide, and that negative carbon cycle feedbacks may have prevented the permanent establishment of large ice sheets earlier than 34 million years ago.

A change in the state of Earth's ocean–climate system occurred at the end of the Eocene (55–34 Myr ago), when the 'greenhouse' climate that had been sustained since the Cretaceous period evolved to the 'icehouse' conditions characterizing the Oligocene epoch through to the present^{1–4}. There is substantial controversy over the history of climate and atmospheric CO₂ concentrations during the greenhouse–icehouse transition. Long-term high-latitude cooling began in the early Eocene (~51 Myr ago), as indicated by an increase in low-resolution oxygen isotope ($\delta^{18}\text{O}$) records from benthic foraminifera^{1,2} and Southern Ocean planktonic foraminifera³ and a decrease in benthic foraminiferal Mg/Ca ratios⁴. Although some studies have concluded that tropical sea surface temperatures also cooled gradually during the Eocene³, most climate proxy records support warm and stable tropical sea surface temperatures throughout the Eocene^{5–8}, with evidence for small cooling steps at 48, 45 and 42 Myr ago from planktonic foraminiferal Mg/Ca data⁸. The earliest Oligocene (~34 Myr ago) is widely accepted as the interval associated with the onset of 'icehouse' conditions. High-resolution reconstructions of seawater $\delta^{18}\text{O}$ show a +1‰ shift that is interpreted as recording a sudden and massive expansion of ice volume, along with the occurrence of ice-rafted debris in the Southern Ocean and a change in clay mineralogy consistent with increased glacial erosion on Antarctica^{4,9–11}. There is stable isotope and lithological evidence for small-scale ice sheets in the Northern Hemisphere during the late Miocene (~10–6 Myr) and the build-up of ice in the Pliocene, beginning around 3 Myr². Others have argued for significantly earlier initiation of ice sheet development using low-resolution proxy records^{12–14}, with some studies concluding substantial ice volume from the late Cretaceous^{15–17}.

Glacial expansion has been linked to a decline in partial pressure of CO₂ on the basis of climate modelling¹⁸, and is supported by evidence for changes in deep water carbonate saturation across the

Eocene–Oligocene boundary¹¹ (grey line in Fig. 1a). Atmospheric CO₂ concentrations declined during the Oligocene, with records from multiple proxies indicating levels of less than 400 parts per million by volume during the last 25 Myr^{19–21}. However, proxy constraints on atmospheric CO₂ levels before 34 Myr ago yield conflicting results. Plant stomata and carbon isotope-based reconstructions support stable greenhouse conditions during the Eocene^{19,20}, whereas boron isotope-based surface water pH reconstructions for the early Eocene indicate highly variable CO₂ concentrations with values ranging from several hundred to a few thousand parts per million by volume²¹.

There are few constraints on climate stability and cryosphere development during the greenhouse–icehouse transition because of the poor resolution of the few existing data sets, particularly for the middle and late Eocene (49–34 Myr ago). Detailed Eocene palaeoceanographic records have been limited to two studies because of diagenetic alteration, hiatuses, coring gaps and lack of stratigraphic constraints, and these studies have inferred large changes in surface water hydrography from planktonic foraminiferal oxygen isotope records in the western Atlantic Ocean²² and the Southern Ocean²³. A recent Ocean Drilling Program (ODP) cruise, ODP Leg 199, successfully recovered a continuous middle Eocene through to early Miocene sedimentary sequence spanning the greenhouse–icehouse transition from the central equatorial Pacific Ocean²⁴. This region is a major locus for production and burial of biogenic carbonate in the modern and glacial ocean. Volumetrically it represents a significant portion of the global ocean, and therefore is used as a 'dipstick' for changes in whole ocean carbon chemistry during the Pleistocene epoch. In the Eocene, the equatorial Pacific would have comprised about two-thirds of all the tropical oceans^{24–26} (Supplementary Fig. 1a), and would have been of even greater importance than during the Pleistocene. This region was dominated

¹Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK. ²Department of Geology and Geochemistry, Stockholm University, Kungstensgatan 45, S-10691, Stockholm, Sweden.

by siliceous deposition until the Eocene–Oligocene boundary at 34 Myr ago, when calcium carbonate nannofossil ooze and chalk replaced radiolarian ooze and radiolarite in less than 300,000 yr (ref. 11), associated with a deepening of the calcite compensation depth CCD, the water depth at which the calcium carbonate rain rate is balanced by the dissolution rate²⁷ of over 1 km (Fig. 1a). The marked CCD deepening has been linked to a rapid growth of the Antarctic ice sheet, based on the coincident 1‰ transient increase in

records of surface and deep water $\delta^{18}\text{O}$ (termed Oi-1) and occurrence of ice-rafted debris in the Southern Ocean⁹ at 34 Myr ago.

The presence/absence of calcium carbonate (0% isopleth) at ODP sites with different palaeodepths (Supplementary Fig. 1) has been used to approximate the basinal CCD in the central equatorial Pacific Ocean (ODP sites 1215, 1217–1221) and the subtropical South Atlantic Ocean (ODP sites 1262–1267). Both records (Fig. 1b) show evidence for basinally synchronous shifts in the CCD beginning at ~46 Myr ago, which is 12 Myr before the Eocene–Oligocene transition. The largest Eocene changes in CCD occurred between about 42 and 38 Myr ago, when both the equatorial Pacific (blue line in Fig. 1b) and South Atlantic CCD (red line in Fig. 1b) deepened and shoaled in unison. The equatorial Pacific CCD was shallow during the late Eocene and underwent a large deepening ~34 Myr ago (blue line in Fig. 1b), whereas the South Atlantic CCD was relatively deep during the late Eocene, and appears to have deepened by only a few hundred metres at ~34 Myr ago (red line in Fig. 1b). Because calcium has an oceanic residence time of 10^6 yr (ref. 27), the middle and late Eocene global CCD variations must record large swings in deep water carbonate saturation, and therefore reflect changes in atmospheric CO_2 levels before 34 Myr ago.

A deep ocean carbonate switch

In order to more accurately resolve CCD changes across the greenhouse–icehouse transition, we developed high-resolution records of calcium carbonate content and mass accumulation rate (MAR) for complete and well-preserved middle and late Eocene sedimentary sequences from equatorial Pacific sites 1218 and 1219. As with the 0% isopleth (calcium carbonate absence) data, these detailed records (Fig. 1c) show that extended intervals of calcium carbonate deposition in the equatorial Pacific Ocean occurred before Oi-1, during the middle and late Eocene. Additionally, the data show rapid, large-amplitude changes, with carbonate values ranging from 0 to 90%, and multiple peaks of carbonate deposition, at intervals of 40,000 and 100,000 yr.

We developed a detailed Eocene CCD reconstruction for the equatorial Pacific (blue line in Fig. 1a) by integrating crustal subsidence curves with carbonate data, averaging the high-resolution MAR for each site over 100,000-yr increments and assuming a linear decrease in calcium carbonate due to dissolution by extrapolating from the difference in MAR observed between Pacific sites 1218 and 1219 (~0.001 g calcite $\text{cm}^{-2} \text{kyr}^{-1}$ per additional metre of water depth near the CCD). Although this method is sensitive to assumptions of dissolution rate and age model, it has the advantage of being insensitive to dilution from non-carbonate components and therefore provides a more accurate alternative means for estimating past CCD. It indicates that the CCD underwent a substantial but gradual deepening in the middle Eocene until ~42 Myr ago (blue line in Fig. 1a), when carbonate accumulation rates rapidly increased (blue line in Fig. 1c). It then rapidly shoaled at ~41 Myr ago, and then decreased to the Eocene–Oligocene boundary with some smaller oscillations during the middle and late Eocene.

The carbonate and accumulation rate data (Fig. 1c) constrain shifts in the middle and late Eocene CCD of 500 m to within 100,000 yr, and of 800 m to less than 200,000 yr, demonstrating that they are comparable in rate, magnitude and structure to those observed across the Eocene–Oligocene boundary¹¹. Two lines of evidence indicate the CCD fluctuations were global. First, the Atlantic (Walvis Ridge) carbonate data contain evidence for CCD fluctuations before 34 Myr ago (red line in Fig. 1b). A MAR-based CCD reconstruction for the Indian Ocean, although discontinuous through the middle and late Eocene, also shows there were several brief CCD excursions during the middle and late Eocene²⁸. In contrast, the CCD record for the equatorial Pacific diverges from other basins^{28,29} across the Eocene–Oligocene transition, with the CCD deepening much more in magnitude and more rapidly in the equatorial Pacific than in other basins, including the North Pacific.

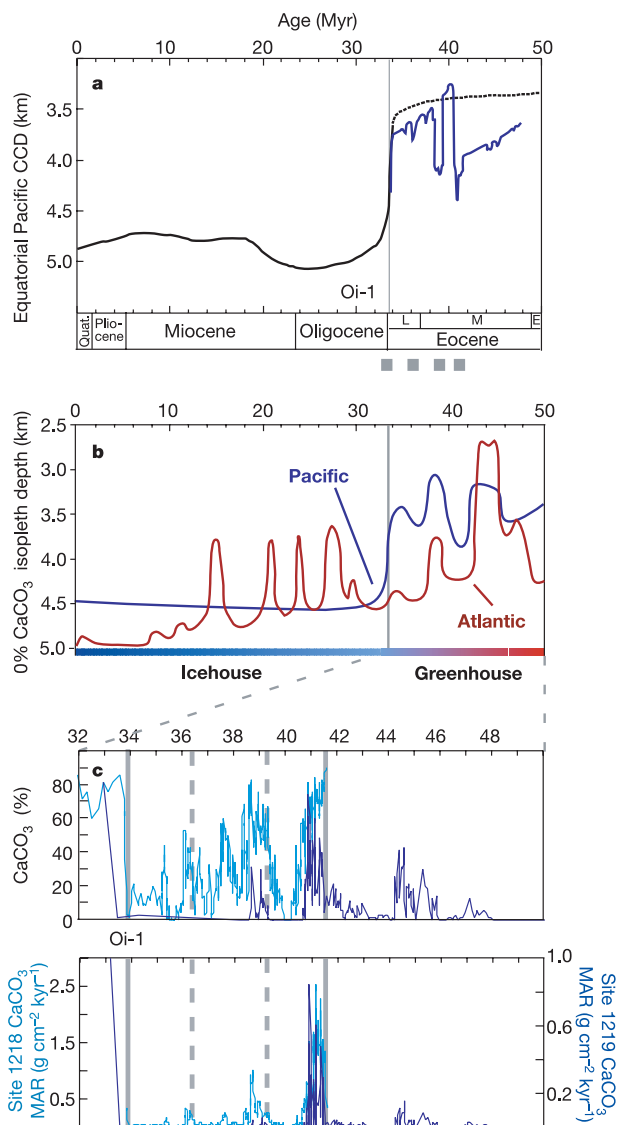


Figure 1 | Records of calcite compensation depth (CCD) and carbonate content for the past 50 Myr. **a**, Reconstructions of CCD based on carbonate mass accumulation rate (MAR) for 0–50 Myr ago. Black line is the low-resolution CCD history based on DSDP sediments²⁵. Eocene sediments (dotted line) were poorly recovered during past drilling in the Pacific Ocean. Blue line is our revised CCD history for the greenhouse–icehouse transition, based on sediments recovered during ODP Leg 199. Grey bars at bottom mark CCD deepening events that coincide with transient benthic $\delta^{18}\text{O}$ increases (Fig. 2; vertical grey lines), and vertical grey line marks Oi-1 glaciation. **b**, Low-resolution reconstruction of 0% calcium carbonate isopleth depth for the equatorial Pacific and subtropical South Atlantic based on presence/absence of carbonate at Leg 199 and 208 sites. **c**, High-resolution records of carbonate content (top) and MAR (bottom) from equatorial Pacific sites 1218 (light blue) and 1219 (dark blue) for 32–50 Myr ago across the greenhouse–icehouse transition. Vertical grey lines coincide with large benthic $\delta^{18}\text{O}$ increases (Fig. 2; vertical grey lines). Solid lines mark major events and dashed lines mark minor events.

Changing ocean circulation and rain ratio influence local CCD and may have driven basal differences observed across the Eocene–Oligocene boundary.

A dynamic middle Eocene cryosphere

In order to study the relationship between changes in CCD and climate, we measured the stable isotope composition of benthic foraminifers and bulk carbonate at Pacific sites 1218 and 1219. The record of benthic $\delta^{18}\text{O}$ (green lines in Fig. 2) displays a similar pattern to that of CCD (blue line in Fig. 1a) across the greenhouse–icehouse transition. When CCD deepened between 48 and 42 Myr ago, benthic $\delta^{18}\text{O}$ values increased (sloping black lines in Fig. 2), reflecting long-term global cooling/higher ice volume. The records also show the same pattern between 38 and 34 Myr ago. On shorter timescales, increased carbonate deposition correlates with high bulk and benthic $\delta^{18}\text{O}$ values. For example, transient positive excursions in $\delta^{18}\text{O}$ (Fig. 3, labelled no. 1; grey lines in Fig. 2) at 42, 39 and 36 Myr ago are contemporaneous with rapid deepenings of the CCD (Fig. 3, blue bar; grey bars in Fig. 1a), as indicated by the widespread appearance of carbonate (grey lines in Fig. 1c). Finer-scale patterns of variability also match up in the records. On millennial timescales we observe a similar correspondence between $\delta^{18}\text{O}$ (red lines in Fig. 3) and carbonate MAR (blue lines in Fig. 3).

The most prominent CCD excursions beginning 42 Myr ago and centred on 41 Myr ago (Fig. 1a) correspond to the most extreme $\delta^{18}\text{O}$

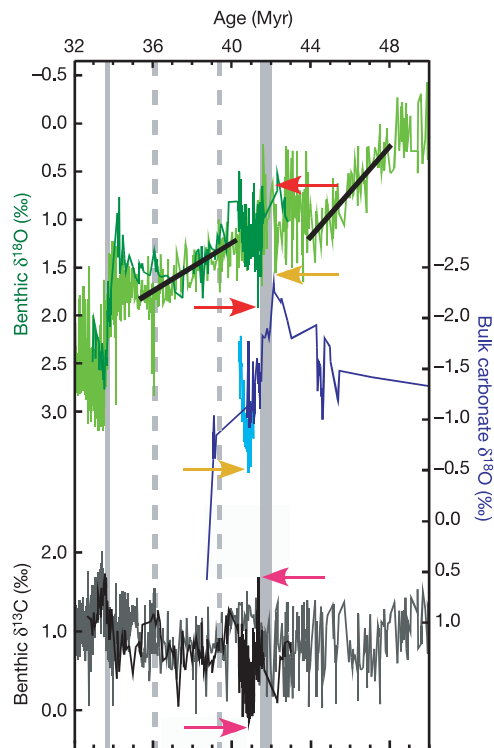


Figure 2 | Stable isotope records across the greenhouse-icehouse transition. Shown are benthic foraminiferal $\delta^{18}\text{O}$ for site 1218 (dark green) and other sites (light green), bulk carbonate $\delta^{18}\text{O}$ for sites 1218 (light blue) and 1219 (dark blue), and benthic foraminiferal $\delta^{13}\text{C}$ for site 1218 (dark grey) and other sites (light grey). Thick vertical grey bar marks the increase in benthic and bulk $\delta^{18}\text{O}$ between 42.0 and 41.5 Myr ago, and is immediately followed by rapid, transient increases in benthic and bulk $\delta^{18}\text{O}$. Thin vertical bars mark transient benthic $\delta^{18}\text{O}$ increases at about 39, 36 and 34 Myr ago. Solid lines mark major events and dashed lines mark minor events. Sloping black lines mark the long-term trend in benthic $\delta^{18}\text{O}$ between 48 and 42 Myr, and between 38 and 34 Myr. Coloured arrows mark the increase in benthic $\delta^{18}\text{O}$ between 42.0 and 41.3 Myr, in bulk $\delta^{18}\text{O}$ between 42.0 and 40.9 Myr, and decrease in benthic $\delta^{13}\text{C}$ between 41.5 and 41.3 Myr. Stable isotope values are relative to the V-PDB standard.

(green lines in Fig. 2) and $\delta^{13}\text{C}$ (grey lines in Fig. 2) values in the Eocene record. This portion of the records, expanded in Fig. 3, shows that benthic $\delta^{18}\text{O}$ increases by 1.2‰, recording some combination of about 6 °C cooling or 120 m of sea level change. The gradual build-up and initial increase in benthic and bulk $\delta^{18}\text{O}$ between 42.0 and 41.5 Myr ago (thick grey line in Fig. 2) precedes the rise in carbonate MAR (grey line in Fig. 1c). The main carbonate pulse at 41.5 Myr ago is associated with a rapid increase in benthic $\delta^{18}\text{O}$ of about 0.9‰ (blue

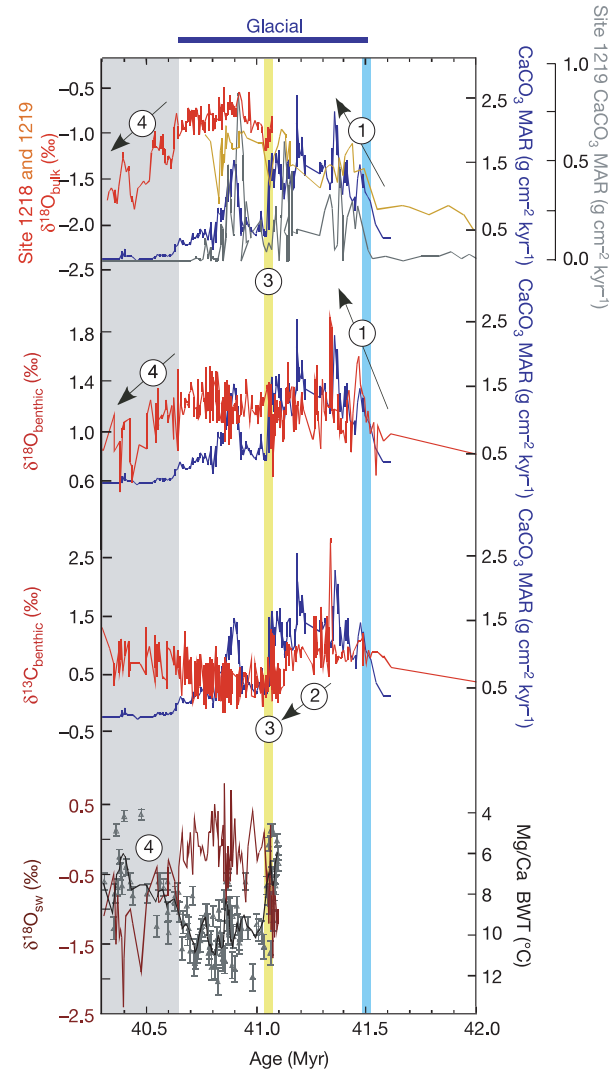


Figure 3 | High-resolution records across the middle Eocene glaciation. Shown are carbonate MAR for sites 1218 (dark blue) and 1219 (grey), carbonate stable isotope ratios, Mg/Ca-based bottom water temperatures (BWT) (grey symbols; black line is smoothed curve fit; error bars are calculated using the average relative s.d. of replicate analyses), and estimates of seawater $\delta^{18}\text{O}$ ($\delta^{18}\text{O}_{\text{sw}}$). Sequential events discussed in the text include: (1) the increase in carbonate preservation marking a deepening of the CCD (blue line) coincident with a change in the $\delta^{18}\text{O}$ of bulk carbonate and benthic foraminifera, signifying growth of continental ice sheets. This CCD deepening is followed by peak benthic $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$ values. (2) A major shift in benthic $\delta^{13}\text{C}$ ratios that we attribute to a carbon budget reorganization in response to glaciation. (3) The 'crash' in carbonate preservation (yellow line) indicating a shoaling of the CCD that is coincident with the minimum in $\delta^{13}\text{C}$ and warming of BWT. This carbonate crash occurs during an interval when bulk carbonate and benthic $\delta^{18}\text{O}$ were increasing, which suggests that glacial expansion was occurring, and is immediately followed by peak BWT and high seawater $\delta^{18}\text{O}$. (4) The decrease in benthic and bulk carbonate $\delta^{18}\text{O}$ (grey interval) indicates deglaciation. The carbonate crash and peak BWT preceded glacial collapse. The estimated change in seawater $\delta^{18}\text{O}$ indicates a sea level change of at least 125 m.

bar in Fig. 3; labelled no. 1). Benthic and bulk $\delta^{18}\text{O}$ continue to increase during several subsequent steps. The decline in carbonate deposition at the end of this event (yellow bar in Fig. 3; labelled no. 3) precedes the decrease in benthic and bulk $\delta^{18}\text{O}$ (grey interval in Fig. 3; labelled no. 4). Benthic $\delta^{18}\text{O}$ decreases within 200,000 yr in several steps, indicating that the transition from cool to warm conditions was as rapid as the transitions at 41.5 Myr ago and across the Eocene–Oligocene boundary^{9,11}.

Several lines of evidence support a large ice volume component to the middle and late Eocene benthic foraminiferal $\delta^{18}\text{O}$ shifts. The timings of CCD deepening and seawater $\delta^{18}\text{O}$ maxima are approximately the same as estimated sea level lowstands^{15,16} in the middle and late Eocene. Combination of the change in Mg/Ca-based temperatures (which is independent of the seawater Mg/Ca model used) with benthic foraminiferal $\delta^{18}\text{O}$ yields a seawater $\delta^{18}\text{O}$ change of over 1.5‰ for the most prominent event (Fig. 3), indicating that there must have been a large ice volume component to the benthic and bulk $\delta^{18}\text{O}$ records. Coincident changes in both benthic and bulk carbonate $\delta^{18}\text{O}$ of up to 1.2‰ (Fig. 3) are consistent with this conclusion.

These constraints on seawater $\delta^{18}\text{O}$ indicate that the largest glaciation was similar in magnitude to Oi-1¹¹. A number of seawater $\delta^{18}\text{O}$ –sea level relationships have been proposed, and even with a conservative estimate (for example, the Quaternary relationship of 0.12‰ per 10 m of sea level), a 1.5‰ change in seawater $\delta^{18}\text{O}$ supports 125 m of sea level variation. If we assume that the isotopic composition of ice was lower during the Eocene (and use a relationship of 0.08‰ per 10 m of sea level)¹¹, the seawater $\delta^{18}\text{O}$ record indicates 190 m of sea level change. Similarly, the shared bulk and benthic $\delta^{18}\text{O}$ shifts support a sea level change of between 100 and 150 m. On the basis of estimates of high-latitude sea surface temperatures of $<12^\circ\text{C}$ (from benthic foraminiferal Mg/Ca in Fig. 3 and other proxy reconstructions^{3,4,14,23}) and reasonable values for Antarctic lapse rates¹⁴ ($6.5\text{--}10^\circ\text{C km}^{-1}$), it is plausible that there was continental storage of ice at high elevations in the Antarctic (for example, in the Transantarctic or Gamburtsev mountains), but not of this magnitude. The $\delta^{18}\text{O}$ values would require there to have been ice storage in the Antarctic during the middle Eocene that was much greater than the modern ice budget for Antarctica¹¹, which is unsupported by any proxy data. Therefore the amount of sea level change necessitates significant storage of ice in the Northern Hemisphere at 42 Myr ago. Early glacial onset in the Northern Hemisphere is supported by the occurrence of ice-rafted dropstones in middle Eocene sediments from Lomonosov ridge in the central Arctic Ocean³⁰, and in the Antarctic by the presence of Eocene glacial sediments in McMurdo Sound¹³ and the transition from smectite to illite/chlorite-rich clay mineral assemblages in middle and late Eocene sediments from Maud Rise and Kerguelen Plateau¹².

Flipping the deep ocean carbonate switch

The initial increase in benthic and bulk $\delta^{18}\text{O}$ between 42.0 and 41.5 Myr ago (thick grey line in Fig. 2) occurs before the build-up of carbonate (grey line in Fig. 1c), indicating that glaciation preceded and may have driven CCD deepening. We hypothesize that the major glaciation 42 Myr ago resulted in a sea level lowering that decreased the area for carbonate deposition on continental shelves and resulted in increased carbonate deposition in the deep basins of the ocean, a mechanism that directly links glaciation with deep ocean carbonate preservation. Global CCD deepening throughout the Cenozoic have been attributed to the build-up of continental ice^{31–33}, which would have lowered sea level, resulting in deepening of the CCD. The erosion of exposed shallow marine carbonates would have also increased riverine inputs of alkalinity, deepening the CCD further (as the response time of carbonate ion in the ocean to changes in alkalinity input is about 10^4 yr; refs 27, 31–33) and increasing seawater $\delta^{13}\text{C}$. We observe that an increase in benthic $\delta^{13}\text{C}$ is coincident with the rise in seawater $\delta^{18}\text{O}$ and carbonate preservation

(Fig. 3; labelled no. 1). Increasing the bicarbonate flux to the oceans through glacial continental weathering could have also contributed to transient CCD deepening³³.

We observe that peak glacial conditions were associated with reduced carbonate accumulation (yellow bar in Fig. 3; labelled no. 3) and warming (Fig. 3; bottom panel), consistent with a rise in atmospheric CO_2 levels during the height of glaciation. Changing riverine fluxes may have been the initial driver for restoring greenhouse conditions. During peak glacial conditions, a slowdown of the hydrologic cycle and reduced silicate weathering (owing to reduced exposure area) would have decreased alkalinity inputs to the oceans³³, thereby forcing a reduction in carbonate deposition, a shoaling of the CCD and a decrease in seawater $\delta^{13}\text{C}$. We find that the pulse of dissolution coincides with the largest benthic $\delta^{13}\text{C}$ decrease in the Eocene (Fig. 3; labelled no. 2): the highest benthic $\delta^{13}\text{C}$ values observed in the global Eocene record decrease by 1.4‰ within less than 500,000 yr to the lowest $\delta^{13}\text{C}$ values (Fig. 3; labelled no. 3) observed in the Eocene deep ocean. This $\delta^{13}\text{C}$ decrease is associated with a rapid increase in benthic $\delta^{18}\text{O}$ (yellow bar in Fig. 3; labelled no. 3). There are other climatic feedbacks that could have also contributed to increased atmospheric CO_2 and decreased seawater $\delta^{13}\text{C}$ during peak glacial conditions. Falling sea level associated with glacial expansion could have resulted in the transfer of organic carbon from exposed continental shelves to the oceans. Cooling-driven acidification might have reduced carbon storage on land and decreased seawater $\delta^{13}\text{C}$, as could changes in rain ratio³⁴. The carbonate ‘crash’ (yellow bar in Fig. 3; labelled no. 3) and increase in Mg/Ca (Fig. 3; bottom panel) are followed by a decrease in the records of benthic and bulk $\delta^{18}\text{O}$ and in the Mg/Ca-based reconstruction of seawater $\delta^{18}\text{O}$ (grey interval in Fig. 3; labelled no. 4), indicating that rising atmospheric CO_2 levels eventually forced the collapse of ice sheets. The records of $\delta^{18}\text{O}$ show that deglaciation occurred within less than 200,000 yr.

Links between cooling and the carbon cycle

These results demonstrate that the greenhouse–icehouse transition was closely coupled to changes in carbon cycling and indicate the presence of extensive ice accumulation before the Eocene–Oligocene boundary, with large fluctuations in ice volume during the middle Eocene. The sequence of changes in our records of sediment and foraminiferal geochemistry is consistent with a sensitive Eocene ocean–climate system with components that were capable of rapid response to forcing. Constraints on global CCD support rapid, transient changes in deep water carbonate saturation and atmospheric CO_2 concentrations during the middle and late Eocene. Both our record of CCD and the boron isotope composition of planktonic foraminifera²¹ support a minimum in atmospheric CO_2 at 42 Myr ago and a maximum at 40 Myr ago.

The largest CCD shifts began about 42 Myr ago and appear to be triggered by glaciation in both hemispheres, and were followed by a set of damped oscillations in ocean chemistry and climate. A series of smaller excursions in carbonate MAR and $\delta^{18}\text{O}$ occur at 39 and 36 Myr ago, presumably indicating smaller glaciations. The subsequent oscillations may reflect the non-steady-state response of the carbon cycle (that is, an attempt to restore a balance between CO_2 sources and sinks) to a perturbation that triggered glaciation 42 Myr ago. The low-resolution boron isotope record has been interpreted as reflecting a reduction in atmospheric CO_2 beginning 45 Myr ago, culminating in a minimum 42 Myr ago (ref. 21). We suggest that gradually decreasing atmospheric CO_2 levels may have resulted in cooling that caused glaciation, possibly associated with increased silicate weathering rates due to early Himalayan uplift. This speculation is supported by the large increase in slope of the seawater strontium isotope curve³⁵ around 42 Myr ago.

The timing of carbonate MAR changes with respect to the $\delta^{18}\text{O}$ record indicate that CCD deepening was initially driven by a shift from shelf to deep-sea deposition associated with falling sea level,

whereas CCD shoaling probably involved a different mechanism. Glaciation would have affected weathering rates and carbon fluxes, probably by affecting riverine inputs of bicarbonate ion or terrestrial productivity, resulting in decreased carbonate ion concentrations and $\delta^{13}\text{C}$ during peak glacial conditions. These changes would have eventually served to shoal the CCD and increase atmospheric CO_2 , a negative feedback that would have ultimately driven deglaciation.

Although glacial expansion across the Eocene–Oligocene boundary was similar in rate and magnitude to the middle Eocene glaciation, the permanent deepening of the CCD 34 Myr ago indicates that negative feedbacks (such as silicate weathering and terrestrial carbon storage) failed to compensate for glacially induced carbon cycle changes. The permanent deepening of the CCD may have been a response to enhanced global siliceous productivity and a change in rain ratio¹¹, which would have caused an increase in carbonate ion concentrations and a decrease in the partial pressure of CO_2 (ref. 34). Alternately, it may reflect a steady-state adjustment to long-term changes in seawater calcium concentrations.

METHODS

Chronology. An orbitally tuned age model to ~46 Myr ago has been developed for sites 1218 and 1219 that integrates foraminiferal, nannofossil, radiolarian, and magnetic polarity datum levels³⁶. The palaeomagnetic reversal stratigraphy for both sites displays a polarity pattern representing chrons C5n through to C20. Magnetic susceptibility, density, pronounced colour and lithology changes have enabled precise correlation of the two sites²⁴. Shipboard age models^{24,37} were used for the other sites.

Carbonate data and CCD. Subsidence histories are shown in Supplementary Fig. 1, and were derived from ODP Initial Report volumes^{24,37}. High-precision measurements of carbonate content were made on a UIC Inc. coulometer with acidification module at Stockholm University. Additional measurements were made using an Analytical Precision 2003 mass spectrometer with a carbonate preparation system at the University of Cambridge, with an estimated precision of better than 2% based on replicate analyses. Shipboard measurements^{24,37} were also used, as were physical property-based estimates for across the Eocene–Oligocene boundary^{11,38}.

Stable isotope ratios. Benthic foraminiferal and bulk carbonate stable isotope ratios were determined on gas source mass spectrometers at the University of Cambridge and Stockholm University. Analytical precision as determined by long-term replicate analyses of a standard is better than 0.08‰. Stable isotope ratios were determined on the benthic foraminifers *Nuttallides truempyi* and *Cibicidoides* spp. that were picked from the > 150 μm size fraction. Stable isotope values have been adjusted to account for species offsets³⁹. These data are compared to a compilation of published data for Atlantic^{2,23}, Indian^{2,23} and Pacific^{2,40} oceans in Fig. 2. All stable isotope ratios are reported in δ notation, in units of ‰ and relative to the V-PDB (carbonate) and V-SMOW (seawater) standards.

Mg/Ca ratios. Mg/Ca ratios were measured in *Oridorsalis umbonatus* and *Cibicidoides* spp. that were picked from the > 150 μm size fraction. Foraminiferal samples were oxidatively cleaned and analysed using a standard protocol described in ref. 8 (and references therein). Initially a set of samples was screened to determine if a reductive cleaning step was necessary to obtain reliable foraminiferal Mg/Ca ratios. Trace element ratios were determined on a Varian Vista ICP-OES using an intensity ratio calibration with an analytical precision as determined by replicate analyses of a foraminiferal standard that is better than 2.8‰, and screened to assess contamination by clays, manganese carbonate overgrowths and other possible phases. Measured Mg/Ca ratios in foraminiferal carbonate can be biased by the presence of clays, detrital grains, adhered carbonates, secondary carbonates, and other contaminant phases, which can be detected using trace element/calcium ratios (including Al/Ca, Fe/Ca, Mn/Ca, Si/Ca, Sr/Ca and Zn/Ca).

Bottom water temperature and seawater $\delta^{18}\text{O}$. Calculated bottom water temperatures use published temperature calibrations⁴¹ and estimates of past seawater Mg/Ca (ref. 42). Error bars are calculated using the average relative standard deviation of replicate analyses. Recent work has shown there may be a saturation state effect on benthic foraminiferal Mg/Ca ratios^{43,44}. Increased deep water carbonate saturation during Eocene glacials may cause us to overestimate glacial temperatures by up to 1–2 °C (ref. 40). Seawater $\delta^{18}\text{O}$ is estimated by combining Mg/Ca-based bottom water temperatures with calcite $\delta^{18}\text{O}$, and thus changes in carbonate ion concentration could potentially bias estimates of seawater $\delta^{18}\text{O}$ by 0.2–0.5‰. Dissolution is unlikely to lead to a significant artefact in these records because benthic foraminifera in the deep ocean can

calcify in waters undersaturated with respect to carbonate ion, and have tests that should be relatively homogenous in trace elements.

Received 31 March; accepted 31 May 2005.

1. Miller, K. G., Fairbanks, R. G. & Mountain, G. S. Tertiary oxygen isotope synthesis, sea level history, and continental margin erosion. *Paleoceanography* **2**, 1–19 (1987).
2. Zachos, J. C., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
3. Zachos, J. C., Stott, L. D. & Lohmann, K. C. Evolution of early Cenozoic marine temperatures. *Paleoceanography* **9**, 353–387 (1994).
4. Lear, C. H., Elderfield, H. & Wilson, P. A. Cenozoic deep-sea temperatures and global ice volumes from Mg/Ca in benthic foraminiferal calcite. *Science* **287**, 269–272 (2000).
5. Adams, C., Lee, D. & Rosen, B. Conflicting isotopic and biotic evidence for tropical sea surface temperatures during the Tertiary. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **77**, 289–313 (1990).
6. Pearson, P. et al. Warm tropical sea surface temperatures in the late Cretaceous and Eocene epochs. *Nature* **413**, 481–485 (2001).
7. Tripathi, A. K. & Zachos, J. Late Eocene tropical sea surface temperatures: A perspective from Panama. *Paleoceanography* **17**, doi:10.1029/2000PA000605 (2002).
8. Tripathi, A. K. et al. Tropical sea-surface temperature reconstruction for the early Paleogene using Mg/Ca ratios of planktonic foraminifera. *Paleoceanography* **18**, doi:10.1029/2003PA000937 (2003).
9. Zachos, J. C., Breza, J. & Wise, S. W. Earliest Oligocene ice-sheet expansion on East Antarctica: Stable isotope and sedimentological data from Kerguelen Plateau. *Geology* **20**, 569–573 (1992).
10. Diester-Haas, L. & Zahn, R. Eocene–Oligocene transition in the Southern Ocean: History of water mass circulation and biological productivity. *Geology* **24**, 163–166 (1996).
11. Coxall, H. K., Wilson, P. A., Palike, H., Lear, C. H. & Backman, J. Rapid stepwise onset of Antarctic glaciation and deeper calcite compensation in the Pacific Ocean. *Nature* **433**, 53–57 (2005).
12. Robert, C. & Kennett, J. P. Paleocene and Eocene kaolinite distribution in the South Atlantic and Southern Ocean: Antarctic climate and paleoceanographic implications. *Mar. Geol.* **103**, 99–110 (1992).
13. Ehrmann, W. Implications of late Eocene to early Miocene clay mineral assemblages in McMurdo Sound (Ross Sea, Antarctica) on paleoclimate and ice dynamics. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **139**, 213–231 (1998).
14. Billups, K. & Schrag, D. P. Application of benthic foraminiferal Mg/Ca ratios to questions of Cenozoic climate change. *Earth Planet. Sci. Lett.* **209**, 181–195 (2003).
15. Browning, J. V., Miller, K. G. & Pak, D. K. Global implications of lower to middle Eocene sequence boundaries on the New Jersey coastal plain—The icehouse cometh. *Geology* **24**, 639–642 (1996).
16. Miller, K. G. et al. Cenozoic global sea-level, sequences, and the New Jersey transect: Results from coastal plain and slope drilling. *Rev. Geophys.* **36**, 569–601 (1998).
17. Miller, K. G. et al. Upper Cretaceous sequences and sea-level history, New Jersey coastal plain. *GSA Bull.* **116**, 368–393 (2004).
18. DeConto, R. M. & Pollard, D. Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric CO_2 . *Nature* **421**, 245–249 (2003).
19. Royer, D. L. et al. Paleobotanical evidence for near present day levels of atmospheric CO_2 during part of the Tertiary. *Science* **292**, 2310–2313 (2001).
20. Freeman, K. H. & Hayes, J. M. Fractionation of carbon isotopes by phytoplankton and estimates of ancient CO_2 levels. *Glob. Biogeochem. Cycles* **6**, 185–198 (1992).
21. Pearson, P. N. & Palmer, M. R. Atmospheric carbon dioxide concentrations over the past 60 million years. *Nature* **406**, 695–699 (2000).
22. Wade, B. S. & Kroon, D. Middle Eocene regional climate instability: Evidence from the western North Atlantic. *Geology* **30**, 1011–1014 (2002).
23. Bohaty, S. M. & Zachos, J. C. Significant Southern Ocean warming event in the late middle Eocene. *Geology* **31**, 1017–1020 (2003).
24. Lyle, M. W. et al. Paleogene equatorial transect. *Proc. ODP Init. Rep.* **199**, 1–87 (2002).
25. van Andel, T. H., Heath, G. R. & Moore, T. C. Cenozoic history and paleoceanography of the central equatorial Pacific. *Geol. Soc. Am. Mem.* **143**, 1–134 (1975).
26. Huber, M. & Caballero, R. Eocene El Niño: Evidence for robust tropical dynamics in the “hothouse”. *Science* **299**, 877–881 (2003).
27. Broecker, W. S. & Peng, T. H. *Tracers in the Sea* (Eldigio, Palisades, 1982).
28. Peterson, L. C. & Backman, J. Late Cenozoic carbonate accumulation and the history of the carbonate compensation depth in the western equatorial Indian Ocean. *Proc. ODP Sci. Res.* **115**, 467–489 (1990).
29. Thunell, R. C. & Corliss, B. H. in *Terminal Eocene Events* (eds Pomeroy, C. & Premoli-Silva, I.) 363–380 (Elsevier, Amsterdam, 1986).
30. Shipboard Scientific Party. Arctic Coring Expedition (ACEX): paleoceanographic and tectonic evolution of the central Arctic Ocean. *IODP Prel. Rep.* **302**, <http://www.ecord.org/exp/acex/302PR.pdf> (2005).

31. Opdyke, B. N. & Wilkinson, B. H. Surface area control of shallow cratonic to deep marine carbonate accumulation. *Paleoceanography* **3**, 685–703 (1998).
32. Delaney, M. L. & Boyle, E. Tertiary paleoceanic chemical variability. *Paleoceanography* **3**, 137–156 (1998).
33. Kump, L. R. & Arthur, M. A. in *Tectonics Uplift and Climate Change* (ed. Ruddiman, W. F.) 399–426 (Plenum, New York, 1997).
34. Archer, D. & Maier-Reimer, E. Effect of deep-sea sedimentary calcite preservation on atmospheric CO₂ concentration. *Nature* **367**, 260–263 (1994).
35. McArthur, J. M., Howarth, R. J. & Bailey, T. R. Strontium isotope stratigraphy; LOWESS Version 3; best fit to the marine Sr-isotope curve for 0–509 Ma and accompanying look-up table for deriving numerical age. *J. Geol.* **109**, 155–170 (2001).
36. Palike, H. *et al.* Astronomical age calibration of Oligocene sediments from ODP Leg 199. *ODP Leg 199 Postcruise Meet. Abstr.* (2003).
37. Zachos, J. C. *et al.* Early Cenozoic extreme climates: Walvis Ridge transect. *Proc. ODP Init. Rep.* **208**, 1–112 (2004).
38. Vanden Berg, M. D. & Jarrard, R. D. Cenozoic mass accumulation rates in the equatorial Pacific based on high-resolution mineralogy of Ocean Drilling Program Leg 199. *Paleoceanography* **19**, doi:10.1029/2003PA000928 (2004).
39. Shackleton, N. J. & Hall, M. A. Oxygen and carbon isotope stratigraphy of Deep Sea Drilling Project Hole 552A: Plio-Pleistocene glacial history. *Init. Rep. DSDP* **81**, 599–610 (1984).
40. Lear, C. H., Rosenthal, Y., Coxall, H. K. & Wilson, P. A. Late Eocene to early Miocene ice sheet dynamics and the global carbon cycle. *Paleoceanography* **19**, doi:10.1029/2004PA001039 (2004).
41. Lear, C., Rosenthal, Y. & Slowey, N. Benthic foraminiferal Mg/Ca-paleothermometry: A revised core-top calibration. *Geochim. Cosmochim. Acta* **66**, 3375–3387 (2002).
42. Wilkinson, B. & Algeo, T. Sedimentary carbonate record of calcium-magnesium cycling. *Am. J. Sci.* **289**, 1158–1194 (1989).
43. Marchitto, T. M., Curry, W. B. & Oppo, D. W. Zinc concentrations in benthic foraminifera reflect seawater chemistry. *Paleoceanography* **15**, 299–306 (2000).
44. Martin, P. A. *et al.* Quaternary deep-sea temperature histories derived from benthic foraminiferal Mg/Ca. *Earth Planet. Sci. Lett.* **198**, 193–209 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Kump for a review of this manuscript, M. Lyle for his efforts and comments, R. Eagle, C. de la Rocha, A. Piotrowski, M. Bickle, S. Crowhurst, R. Alley, T. van Andel and N. Shackleton for discussions of this work, M. Hall, J. Rolfe, L. Booth, M. Greaves, S. Farquhar and C. Sindrey for their technical help, and the Leg 199 Scientific Party for their efforts. This research used samples and data provided by the Ocean Drilling Program (ODP). This work was supported by the British Council through a Marshall Sherfield Postdoctoral Fellowship (A.T.), by NERC and the Comer Foundation (A.T. and H.E.), and by the US Science Support Program (A.T.). J.B. was supported by the Swedish Research Council.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.T. (atri02@esc.cam.ac.uk).

Neural crest origins of the neck and shoulder

Toshiyuki Matsuoka^{1,2*}, Per E. Ahlberg^{4*}, Nicoletta Kessar¹, Palma Iannarelli¹, Ulla Dennehy¹, William D. Richardson^{1,3}, Andrew P. McMahon⁵ & Georgy Koentges^{1,2,3}

The neck and shoulder region of vertebrates has undergone a complex evolutionary history. To identify its underlying mechanisms we map the destinations of embryonic neural crest and mesodermal stem cells using Cre-recombinase-mediated transgenesis. The single-cell resolution of this genetic labelling reveals cryptic cell boundaries traversing the seemingly homogeneous skeleton of the neck and shoulders. Within this assembly of bones and muscles we discern a precise code of connectivity that mesenchymal stem cells of both neural crest and mesodermal origin obey as they form muscle scaffolds. The neural crest anchors the head onto the anterior lining of the shoulder girdle, while a *Hox*-gene-controlled mesoderm links trunk muscles to the posterior neck and shoulder skeleton. The skeleton that we identify as neural crest-derived is specifically affected in human Klippel-Feil syndrome, Sprengel's deformity and Arnold-Chiari I/II malformation, providing insights into their likely aetiology. We identify genes involved in the cellular modularity of the neck and shoulder skeleton and propose a new method for determining skeletal homologies that is based on muscle attachments. This has allowed us to trace the whereabouts of the cleithrum, the major shoulder bone of extinct land vertebrate ancestors, which seems to survive as the scapular spine in living mammals.

The vertebrate neck has undergone a drastic evolutionary transformation from an immobile bony bridge between head and shoulder in early vertebrates with paired fins¹ to a mobile system of muscle scaffolds interconnecting the head and shoulders in early jaw-bearing fish such as placoderms². These scaffolds have retained a structure and function that has been remarkably conserved in jaw opening and head mobility ever since³. The fundamental changes in the skeleton of the neck and shoulders reflect evolving embryonic differentiation processes of mesenchymal stem cells: from bone to muscle connective tissues and cartilage. These have defied mechanistic analysis because the detection of fate changes in homologous cell populations requires experimental long-term lineage labelling which has so far only been possible in the chick⁴, a species with a highly modified neck architecture⁵. Gills and most of the head skeleton are derived from the embryonic cranial neural crest⁴, whereas the limb skeleton is derived from trunk mesoderm⁶. The neural crest and mesoderm do not provide obvious landmarks for their respective boundaries in the intervening neck transition zone: the cranial neural crest is not segmentally deployed in this post-otic region (behind rhombomere 5)⁷ and limb lateral plate mesoderm does not seem to pattern the shoulder girdle proper⁸. With post-otic neural crest (PONC) and paraxial (somitic) mesoderm as candidate components, the neck between the ear (otic) capsule of the head and the trunk forelimbs has remained an uncharted embryonic territory.

Bone formation versus muscle scaffolds

Neural-crest and mesodermal cells seem to differ in the way in which they form bones: neural crest forms dermal and endochondral bones in the head whereas mesoderm forms endochondral skeleton in the trunk. So far no evidence for mesoderm-derived dermal bones has been produced. The shoulder girdle and neck between head and

limbs contains both dermal and endochondral bones. All previous investigations into the evolution of this region have therefore assumed this dermal–endochondral distinction to be a safe indicator for bone origins and homologies: Accordingly, all dermal bones in the post-otic region are considered to be derived exclusively from the neural crest, whereas all endochondral bones are considered mesodermal^{9,10}. The validity of this widely held ‘ossification model’ has remained untested in the neck of any living vertebrate. Indeed, in apparent contradiction of it, a current view holds the posterior boundary of neural-crest-derived skeleton to be the parietal (or frontal) bone of the skull^{11,12}: no neural-crest-derived skeleton behind the ear capsule has yet been identified.

Comparative neck anatomy in living jawed vertebrates challenges the likelihood of the prevailing ‘ossification model 1’. We note that the pattern of neck muscles (red in Fig. 1) is far more conserved than the ossification modes of the shoulder bones to which these muscles are attached (Fig. 1; pale grey regions are dermal, dark grey regions are endochondral bones). This poses a serious problem for muscle homologies: in all cranial and trunk regions of the vertebrate body so far examined the embryonic cellular origin of muscle connective tissues and their respective skeletal attachment regions are identical^{13,14}. This implies that if attachment regions change in their cellular origins and ossification type, their coordinated muscle connective tissues also change in their composition. This would force us to reject the homology of all neck musculature in jawed vertebrates, although it has a highly similar and complex connectivity pattern (red in Fig. 1, ref. 5). A similar problem is posed by the composition of the clavicular bone itself (box 3 in Fig. 1). It is dermal in fish and amphibians, and of mixed dermal plus endochondral composition in mammals^{15,16}. If the ‘ossification model’ holds, we would have to refute the homology of a bone that has changed its histogenesis but

¹Wolfson Institute for Biomedical Research, ²Laboratory of Functional Genomics, ³Department of Biology, University College London, Gower Street, London WC1E 6BT, UK.

⁴Subdepartment of Evolutionary Organismal Biology, Department of Physiology and Developmental Biology, Uppsala University, Norbyvägen 18 A, 752 36 Uppsala, Sweden.

⁵Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

*These authors contributed equally to this work.

not its position inside the head–neck assembly over more than 400 Myr. In this light a competing ‘muscle scaffold model 2’ can be considered, according to which bones ‘morph’ around a highly constrained muscle attachment scaffold¹³. Cell population boundaries remain stable and are the structural basis for conserved muscle scaffolds, but the differentiation of PONC and mesodermal cells into bone, cartilage or muscle connective tissues varies because of changes in signalling pathways. This implies that cell population boundaries are cryptic; they do not coincide with bone sutures but with muscle attachment sites.

By using a recombinase-mediated genetic lineage labelling strategy in transgenic mice we can now discriminate between these two models. We map neck neural crest and mesoderm with single-cell resolution onto muscular (connective tissue) attachment points and skeletal structures of a given (dermal versus endochondral) ossification type (Fig. 2). The two models make mutually exclusive predictions for shoulder girdle origins: if model 1 is correct, the anterior scapular spine would be mesodermal because it is endochondral (left part of box 1 in Fig. 1); if model 2 is correct, the same skeletal region (box 1 in Fig. 1) would be neural crest because it serves as the

attachment region for the trapezius muscle (T in Fig. 1) with expected neural-crest connective tissues.

Here we reveal a cryptic neural crest–mesoderm boundary inside the neck and shoulder girdle skeleton, which ignores traditional skeletal landmarks or (endochondral versus dermal) ossification types and thus invalidates the traditional ‘ossification model’. Instead, cellular distributions of neural crest and mesoderm correspond precisely to muscle attachment scaffolds to the shoulder girdle, corroborating the non-intuitive ‘scaffold model’. This finding illuminates the aetiology of various hitherto poorly understood congenital diseases in humans that are co-extensive with neural-crest-derived shoulder structures. By using the ‘scaffold model’ as a new arbiter of bone homologies, palaeontology can date fate changes of common precursor populations in fossils. This reveals an unexpected evolutionary directionality in underlying fate decisions of mesenchymal stem cells that originate from mesoderm and neural crest.

Cryptic neural crest in neck and shoulders

The key problem we wish to address is the full distribution of skeletal PONC. By using *Wnt-1* (ref. 17) and *Sox-10*–Cre-recombinase-mediated fate mapping (Fig. 2a) we ask three questions. First, can we find evidence that PONC forms endochondral bones? This determines whether either the ‘ossification’ or the ‘scaffold model’ are applicable to the neck region (Fig. 2b). Second, is the entire dermal skeleton behind the otic capsule derived from the neural crest, or is some of it mesodermal (Fig. 2a, b)? This will test the validity of the ‘ossification model’ in the only species that is currently accessible to high-resolution lineage mapping: the mouse. Third, does the distribution of neural crest and mesoderm correlate with muscle attachment points or with ossification types in the neck and shoulder skeleton? This will distinguish the explanatory value of the ‘ossification model’ from that of the ‘scaffold model’ because each model makes non-overlapping predictions about anterior shoulder girdle origins.

Neural crest proves to have an unexpectedly pervasive role in the mouse neck region, forming bone, cartilage and muscle connective tissue within two domains: first, an external, essentially tubular domain dominated by pharyngeal arch muscles that extends from the head to the entire ancestral shoulder girdle and incorporates its anterior part; second, a ventral internal domain comprising internal pharynx and larynx constrictors, tongue muscles, thyroid, cricoid and arytenoid cartilages and their respective muscle attachments at the oesophageal entry (Fig. 2b). Ensheathed between these two tubular domains lies a mesodermal domain centred on the somite-derived vertebral column, which reaches forward to the occipital region of the head (oc in Fig. 3).

The largest component of the external crest domain is the trapezius muscle and its attachment regions (tra in Fig. 3a–c). This is a branchial muscle, innervated by the accessory nerve X/XI with a position that has remained remarkably conserved in all jawed vertebrates⁵. The cranial neural crest connectivity code revealed by our previous chick–quail work¹³ led us to predict that the connective tissue of the trapezius and all its post-otic attachments should be formed from PONC—even if they are endochondral. This indeed proves to be so. At the anterodorsal end, a patch of LacZ⁺/GFP⁺ PONC endoskeleton forms the occipital protuberance inside otherwise mesodermal (occipital) territory; that is, the trapezius attachment point in the skull (Fig. 3a). The labelling extends to the attached nuchal ligament (nl in Fig. 3a), the trapezius muscle connective tissues (Fig. 3b) as well as their respective endoskeletal attachment region on the entire anterior margin and inside the scapular spine (pc, Fig. 4a, b), the coracoid, acromion (Fig. 4c, box 2 in Fig. 1) and also the periosteal muscle attachment caps on spinous processes of all cervical vertebrae (sp of C₄ in Fig. 3c). Thus, the posterior neural crest boundary is found inside the shoulder girdle endoskeleton, where it forms an anterior attachment region of branchial muscles.

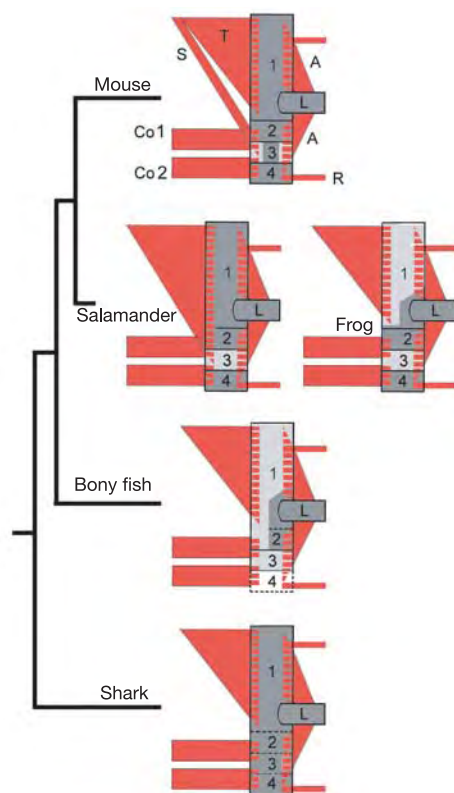


Figure 1 | Conservation of muscle scaffolds and diversity of shoulder ossification patterns. Highly conserved neck muscle scaffolds (red) attach (hatched areas) on a shoulder skeleton (boxes 1–4) that displays variable dermal (light grey) and endochondral (dark grey) ossification type. Attachment regions (hatching) of the gnathostome trapezius muscle (T)^{5,50} are endochondral in sharks, salamanders⁵ and all amniotes but are dermal in fish and frogs. A, limb muscles; Co1 and Co2, coraco-branchialis and coraco-hyoideus; L, limb skeleton; R, trunk muscles; S, sterno-cleido-mastoid. In the shoulder skeleton, box 1 is the dorsal cleithrum (dermal) in bony fish (*Polyodon*⁵ or *Amia*³⁵) and frog (*Rana*³⁶) and the scapular region (endochondral) in salamander⁵, mouse and living amniotes; box 2 is the acromio-coracoid (endochondral); box 3 is the clavicular region (dermal/dermal + endochondral), although in sharks⁵ bone is absent and its space is taken by part of the scapulo-coracoid (stippled); and box 4 is the sternal region, comprising the sternum (endochondral) or connective tissue (bony fish).

PONC thus generates extensive areas of endochondral ossification in the shoulder girdle and cervical vertebral column, contradicting the traditional notion that these regions are wholly mesodermal^{4,11}. However, the crest contributions are morphologically cryptic: their only visible anatomical landmarks are the branchial muscle attachments. Tracing the posterior PONC boundary more ventrally we investigate the sterno-cleido-mastoid muscle (N. XI; scm in Fig. 5a, b, c, e). This originates on the post-otic mastoid process of the skull (m in Fig. 5a, b) and attaches onto the endoskeletal anterior sternum (st in Fig. 5a, e) as well as along the anterior margin of the dermal clavicle (cl in Fig. 5a, c, d). We find green LacZ⁺/GFP⁺ PONC cells inside all these attachment sites. Our genetic labelling provides first detailed insights into cellular architecture and origins of the clavicle (cl, black box in Fig. 5). The clavicle bone forms from an anterior (buccal) dermal ossification centre (Fig. 5d) and a posterior dermal ossification centre (cld, Fig. 5g), which later fuse and surround a cartilaginous core in mammals (Fig. 5c, g)^{15,16}. The anterior dermal ossification is derived purely from PONC (green box, Fig. 5d). More medially, attachment regions of the (sterno-cleido-mastoid (scm) and the fascial sling of omohyoid (ohs) branchial muscles as well as most of the cartilaginous core of the clavicle are also derived from neural crest (Fig. 5c). Thus, as inside the scapula, PONC inside the clavicle gives rise to endochondral bone.

The ventral shoulder girdle carries a series of muscles that connect

it to ventral branchial elements (Box 4, Co1 and Co2 in Figs 1 and 7): M. omohyoideus (connecting the anterior scapula next to the coracoid, and the internal clavicle, with the hyoid; oh in Figs 4c and 5a), M. sternohyoideus (connecting the manubrium sterni and clavicle to the hyoid; sh in Fig. 4d) and M. sternothyroideus (sth in Fig. 4d), connecting the manubrium sterni (mst) with the thyroid cartilage (t in Fig. 5a). These are homologues of the coracobrachial muscles (Co1 and Co2), which are present in all jawed vertebrates and effect rapid jaw opening and retraction of the branchial skeleton^{3,5}. As these are in origin branchial muscles (innervated by cranial nerves⁵) we proposed that their connective tissue would be crest derived; this is indeed so. Swallowing in all jawed vertebrates requires internal pharynx and larynx constrictors (const. phar. in Fig. 6g), which constitute the internal tubular crest domain outlined above. These are connected to the mesodermal ventral neck vertebrae through the pre-vertebral ligaments (Fig. 6e) as well as to the mesodermal cranial base (the so-called clivus) anterior to the foramen magnum through the pharyngobasilar fascia (Fig. 6g). Their branchial innervation by N.IX/X. suggests that their attachment regions area also PONC-derived in all vertebrates. Despite being at odds with the commonly held notion that 'chordal' cranial base is entirely mesodermal^{4,11}, our PONC-labelled mice show constrictor muscle attachment points of neural crest origin (attp in Fig. 6g) focally inserted into the otherwise mesodermal endochondral cranial

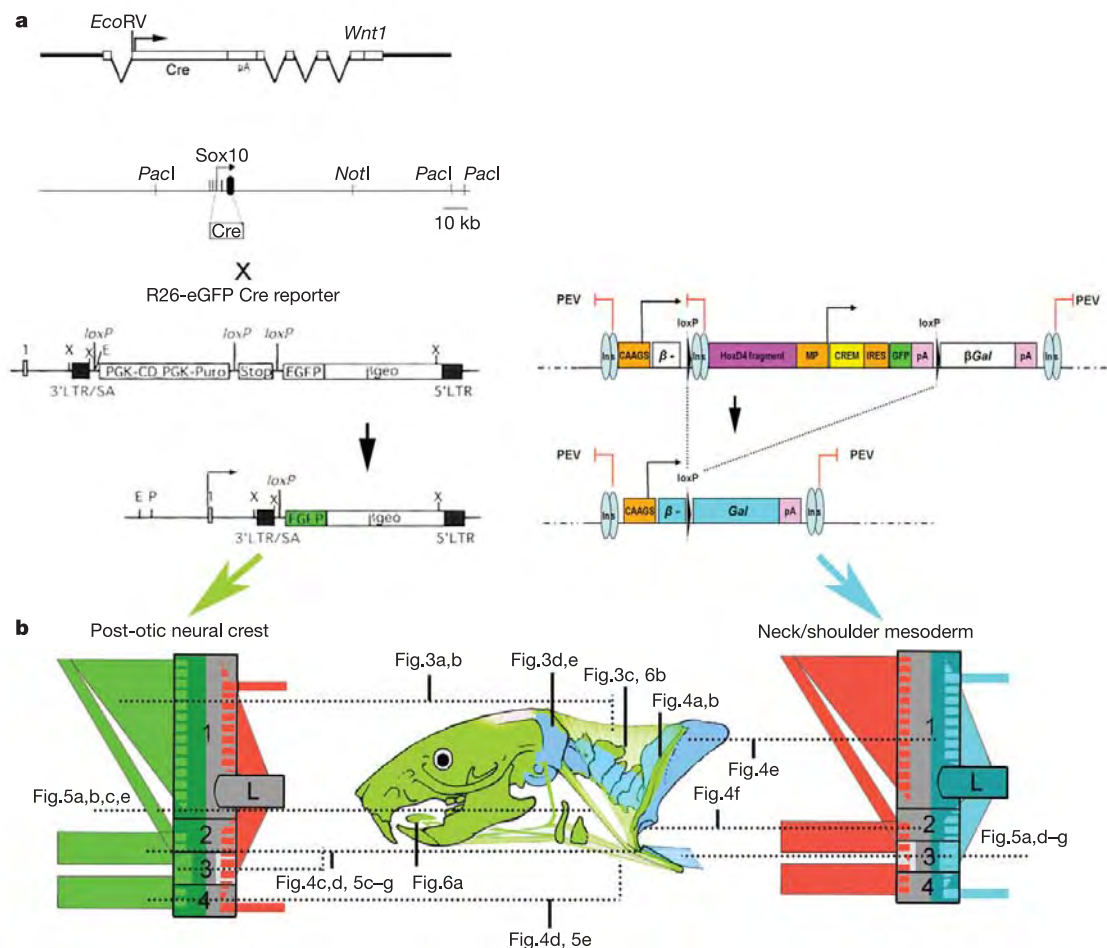


Figure 2 | Genetic lineage labelling of PONC and somitic mesoderm in the neck. **a**, Transgenic mice carrying a *Wnt1*-Cre construct (ref. 17) (top left) or a 170-kilobase *Sox10*-Cre construct (bottom left) were crossed into R26-LacZ⁴⁵ and R26-eGFP (ref. 46) reporters, to permanently label all pre- and post-migratory post-otic/neck neural crest (green). Right: a self-excising Cre recombinase leads to the enhancer-dependent reconstitution of

a β -Gal reading frame (blue) in the complementary (*HoxD4*⁺) mesodermal stem cell population of somite 5 and posterior (blue). Technical details are given in Methods and Supplementary Methods S1. **b**, Anatomical guide through mouse neck and shoulder skeleton, with keys to diagrams and micrographs.

base (mes in Fig. 6g). More posteriorly, thyroid, cricoid and arytenoid cartilages and their respective muscle attachments at the oesophageal entry are also derived from neural crest, with tracheal cartilages demarcating the anterior mesoderm boundary (data not shown). The entire intrinsic tongue musculature that is attached to crest-derived branchial skeleton has crest-derived connective tissue, despite being innervated by N.XII and cervical spinal nerves (Fig. 6a). This demonstrates that motor innervation alone cannot serve as a reliable indicator of connective-tissue origins of embryonic muscle, but skeleton-muscular connectivity can. These genetic PONC-labelling experiments show that PONC in mouse behaves in the same way as the pre-otic crest studied in our previous experiments¹³. Crest cells form not only the connective tissue of the muscle but all its attachment points, irrespective of how these attachment points ossify, whether endochondrally or dermally or whether—as inside the tongue (Fig. 6a)—they do not ossify. This unveils a pervasive but anatomically cryptic ‘muscle scaffold system’ that is sharply defined at the single-cell level.

Rules of engagement in the mesodermal neck

In addition to branchial muscles at its anterior margin, the shoulder

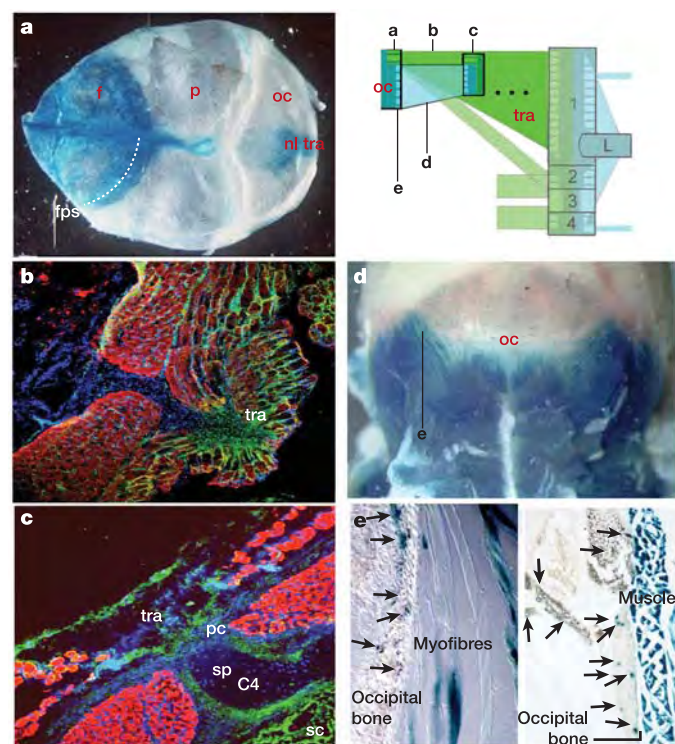


Figure 3 | Neural crest and somitic mesodermal origins of the neck. Green areas in the diagram are neural-crest-derived connective tissues; blue areas are mesodermal (somitic) connective tissues and muscle fibres. **a–c**, Neural-crest-derived (nl, nuchal ligament) attachment regions of the trapezius (tra) onto the occipital skull (oc in **a**) and perichondral (pc) sheaths around spinous processes (sp) of cervical vertebrae (C₄) are LacZ⁺ in Wnt-1 transgenic (**a**) and GFP⁺ (green) in Sox10 transgenic (**b**, **c**) mice. As an internal control the dorsal spinal cord (sc) is GFP⁺ in Sox10-Cre-reporter crosses. Here and in all following figures, nuclear 4,6-diamidino-2-phenylindole (DAPI) stain is blue, myosin heavy-chain immunoreactivity of muscle fibres is red, and GFP⁺ neural crest cells are green. The fronto-parietal suture (dotted line in **a**, fps) between frontal (f) and parietal (p) bones does not correspond to any cellular boundary of LacZ⁺ neural crest (compare with Supplementary Methods S1). **d**, **e**, The mesodermal connectivity system. HoxD4–LacZ⁺ mesodermal occipital bone (oc in diagram and **d**, arrows in **e**) is connected to neck vertebrae (box in diagram; dotted lines represent other vertebrae omitted) directly through striated myofibres (**e**) of the same genetic (HoxD4–LacZ⁺) axial identity.

girdle in all jawed vertebrates serves as an attachment for mesodermal trunk and limb muscles (with spinal motor neuron innervation and connective tissues derived from mesoderm) at its posterior margin⁵. Mesodermal trunk muscles also attach to the occipital head (oc in Fig. 3)^{11,14}. Although the mesodermal (somitic) origins of vertebrae and occiput have long been established, the rules of connectivity between muscles and bony elements have remained unclear. This prompted us to test whether skeletal attachment sites of muscles with mesodermal connective tissues are also mesodermal and of the same (*Hox* gene) axial identity. As axial *Hox* gene expression boundaries sometimes do not coincide with somitic boundaries¹⁸, a genetic approach is required (Fig. 2a, right side and Supplementary Methods S1).

An analysis of HoxD4–CREM–LacZ transgenic mice shows that that posterior margin of the scapular spine (right part of Box 1 in Figs 1 and 7; blue in Figs 2b and 4e, f) and also that the entire scapular blade is of somitic (LacZ⁺) origin at places where mesodermal

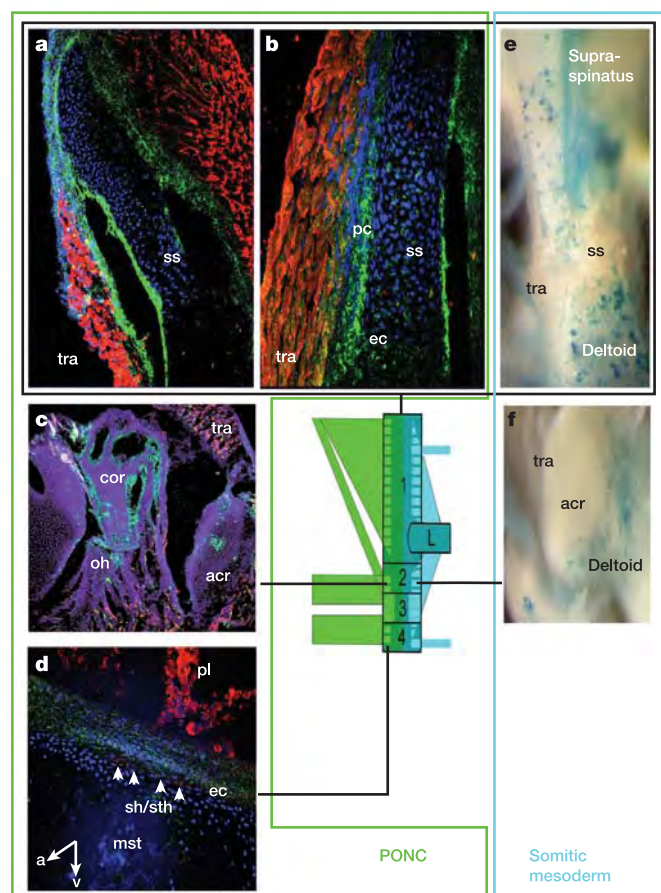


Figure 4 | Dual neural crest and mesodermal origins of the endochondral shoulder girdle. Green box, PONC GFP⁺ green components (**a**, **b**); blue box, mesodermal LacZ⁺ blue components (**e**) of the scapular spine (box 1 in diagram; **a**, **b**, **e**), the acromio-coracoid (box 2; **c**, **f**) and sternum (box 4; **d**). Inside the scapular spine (ss in **a**, **b**) GFP⁺ neural crest cells are found to form endochondral skeleton (ec) and perichondrium (pc) exactly at the places where the trapezius (tra) carrying crest connective tissue is attached. Conversely, blue mesodermal muscle fibres are attached to punctate areas of the posterior scapular spine (**e**) and acromion (**f**). Note that areas of trapezius attachment in mice with mesodermal labelling are white (that is, unlabelled), showing that neural crest and mesodermal muscle attachment systems do not mix with each other. Coraco-brachial sterno-hyoideus/-thyroideus (sh/sth) muscle fibres in **d** are inserted by the endochondral PONC component (white arrowheads) onto the mesodermal sternum (mst). a, anterior; acr, acromion; oh, omo-hyoideus muscle; pl, pleural muscles; v, ventral.

muscles (supraspinatus and deltoid; Fig. 4e, f) are attached. Similar to the scapular spine, the more ventral coracoid and acromion (Box 2 in Figs 1 and 7; Fig. 4c, f) are also of dual neural-crest-mesoderm origin with attachment regions corresponding to muscle connective tissue origins. More ventrally, the clavicle reveals an as yet unrecognized potential of somitic mesoderm to differentiate into dermal bone (Box 3 in Figs 1 and 7; cld in Fig. 5g). According to the commonly held 'ossification model', all dermal clavicular parts (that is, both the anterior and the posterior parts) are expected to be derived from neural crest^{9,10}. We find this to be an incorrect prediction for the clavicle. The posterior margin of the clavicle in *HoxD4*-LacZ transgenic mice is LacZ⁺ at all dermal (cld in Fig. 5g) and endochondral (cle in Fig. 5g) attachment sites of LacZ⁺—that is, mesodermal pectoral and deltoid—muscle fibres (pe and de in Fig. 5f, g).

This shows not only that the clavicle itself is a neural-crest-mesodermal interface but also that postcranial mesoderm gives rise to dermal skeletal elements. It was well known that the posterior

dermal clavicle ossifies independently from the anterior dermal ossification centre, but its separate (mesodermal) origin was unknown^{15,16}. This is the first experimental precedent for interpreting other trunk dermal armour plates among fossil and extant vertebrates as mesodermal. On the basis of our findings it is conceivable that the posterior dermal clavicle is the last remnant of a more widespread body armour of mesodermal origin. More ventrally, the sternum (Box 4 in Figs 1 and 7) is of mesodermal origin at its core and posterior margin, in line with the connectivities described above (Fig. 5e, unlabelled st area, and data not shown), and its anterior margin (manubrium sterni, mst in Fig. 4d and st in Fig. 5a, e) harbours crest-derived endochondral attachment points for coracobrachial musculature (Co2, box4 in Fig. 1, 4, 7, arrowheads in Fig. 4d).

For the occipital head region (Fig. 3d, e), examination of *HoxD4* transgenic mice shows that blue (*HoxD4*-LacZ⁺) muscle fibres are focally and directly attached to *HoxD4*⁺ skeletal regions (arrows in Fig. 3e). This shows by genetic means that somitic mesoderm strictly

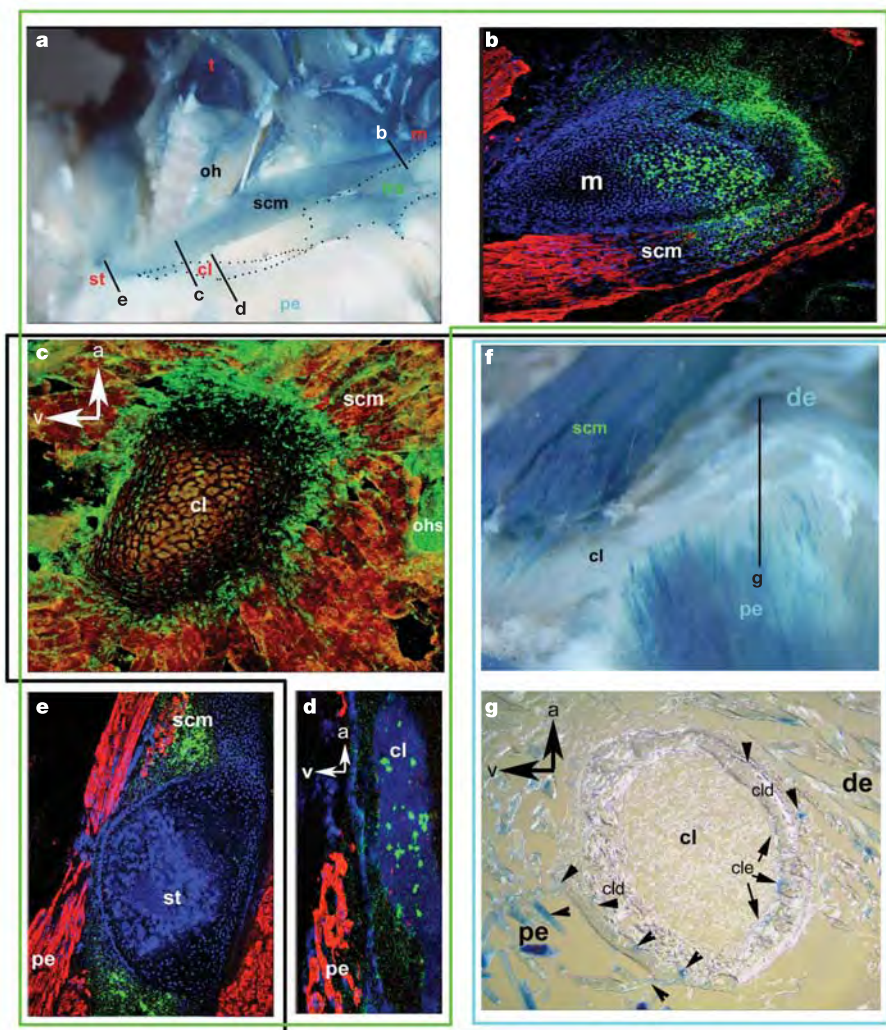


Figure 5 | The dual architecture of the clavicle: cell population boundaries coincide with muscle attachments and not with ossification modes. The black box surrounds data for the clavicle as a bone. **a–e**, PONC-derived parts (surrounded by green line); **f, g**, mesodermal parts (surrounded by blue line). **a**, Anatomical key to other panels. **b–e**, PONC connective tissue of the sternocleidomastoid (scm) attaches onto the mastoid process (m) of the post-otic skull (**b**) and reaches down to the clavicle (cl), which is endochondral (in **c**) and dermal (in **d**) anteriorly, and onto an endoskeletal PONC patch of the sternum (st in **e**). Contiguous DAPI staining of bone

matrix in **d** shows that no other unlabelled (that is, mesodermal) cells are present inside the anterior clavicle. **f, g**, Complementary to this, LacZ⁺ mesodermal myocytes of the pectoral (pe) and deltoid (de) muscles are attached exclusively onto the posterior dermal (cld) and endochondral (cle) clavicle margin (arrowheads in **g**). The anterior clavicle is unlabelled in the transgenic mouse with mesodermal labelling (white in **f**): LacZ⁺ mesodermal cells are confined to mesodermal attachment points. Ohs, fascial sling of the M. omo-hyoideus (oh) onto the clavicle; a, anterior; v, ventral.

obeys a *Hox*-mediated genetic connectivity code directly comparable to that of the neural crest. Although the trapezius and sternocleidomastoid muscle also receives all its muscle fibres from *HoxD4*⁺ somitic mesoderm (Fig. 3d; Fig. 5f, blue scm) these fibres are connected to skeletal elements only through PONC-derived connective tissues (Figs 3b, c and 5a–c, e). In contrast, *HoxD4*⁺ mesodermal muscle fibres of occipital, pectoral or deltoid muscles connect directly to mesodermal skeletal structures, without mediation through mesodermal connective tissues (Figs 3e, 4e, f, 5g). Post-otic somitic mesoderm is similarly capable of forming dermal and endoskeletal bone as well as connective tissue, a fact that invalidates the universal use of the ‘ossification model’ in comparative morphology or palaeontology. Instead, we find that the muscular ‘scaffold model’ more appropriately describes the dual composition and complex connectivity pattern of the neck and shoulder girdle.

Neural crest and human neck pathology

The flexibility of shoulder ossification types inside a highly constrained (trapezius) muscle scaffold as observed in comparative neck anatomy (T in Fig. 1) might reflect an innate flexibility of PONC cells to respond to osteogenic stimuli that, in turn, might render them prone to pathological malformation. The dual origin of neck and shoulder structures would also make it likely to find modular—that is, PONC-specific versus mesoderm-specific—pathological phenotypes. We therefore searched for human congenital syndromes with

a tissue distribution that precisely corresponds to the PONC-population territory but displays pathological differentiation of PONC cells only: crest connective tissues would adopt new cartilaginous or osseous fates and vice versa (Fig. 6). The complex distribution of PONC-derived structures and the strict muscular connectivity rules yield safe criteria for discriminating patterning from differentiation defects. Several hitherto poorly understood human syndromes precisely match the profile of a PONC syndrome and permit first insights into their common cellular aetiology. Klippel–Feil disease (Fig. 6d, f)¹⁹, Sprengel’s deformity (Fig. 6d)²⁰, cleidocranial dysplasia (Fig. 6c)²¹, Arnold–Chiari I/II malformation (Fig. 6h)²² and ‘cri-du-chat’ syndrome (data not shown)²³ are all characterized by co-occurrence of pharyngeal/laryngeal, (sub-)occipital, cervical and shoulder dysmorphologies and swallowing problems. They also share a spina bifida occulta (Fig. 6c, d) that is unusually confined to the cervical region normally occupied by the trapezius muscle (tra in Figs 3c and 6b). In Sprengel’s deformity a large fibrous, sometimes endochondral, so-called omo-vertebral bone replaces all dorsal neural-crest-derived endochondral elements of occipital region, cervical spinous processes, spina scapulae and trapezius²⁰ inside the PONC trapezius territory (Fig. 6d). On this basis we identify Sprengel’s deformity, which is one of the phenotypic facets of Klippel–Feil syndrome, as primarily affecting PONC fate choices and not cervical segmentation as is currently thought¹⁹. Moreover, the cervical hypomobility of Klippel–Feil patients can also be understood as caused by defects in PONC fate choices: ectopic

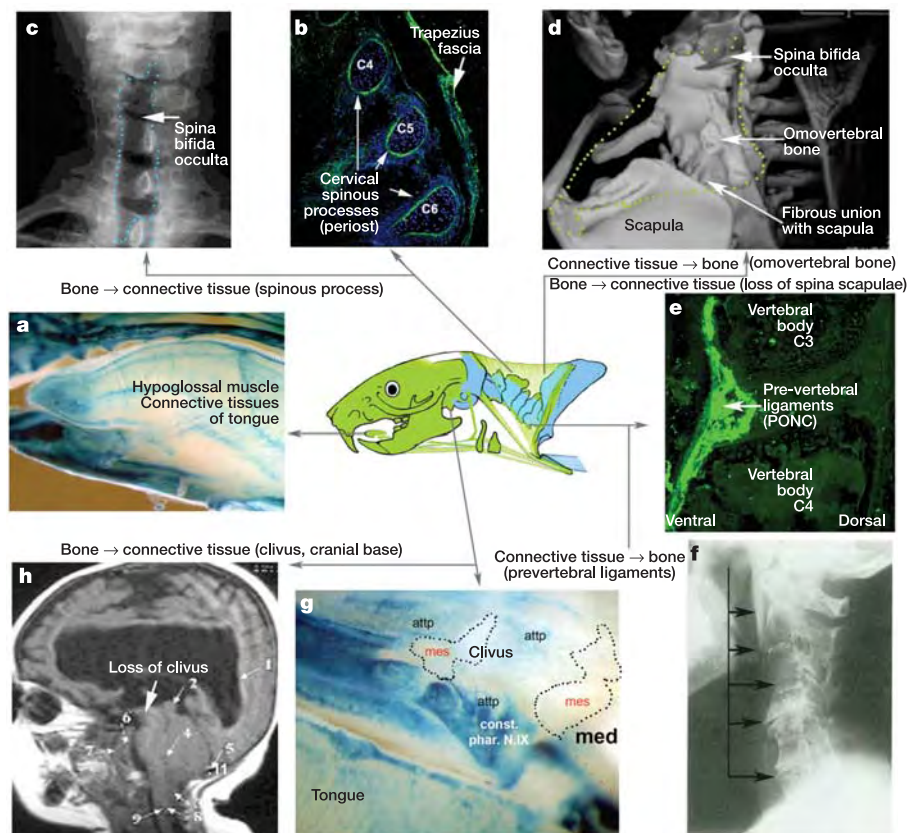


Figure 6 | Pathological flexibility of PONC differentiation. All tongue-muscle connective tissues are entirely derived from neural crest (blue areas in **a**), explaining enlarged tongues (in patients with trisomies) as neurocristopathic. Left: changes in neural crest fate specification from bone periosteum (**b**) into connective tissue can explain localized cervical defects in cleidocranial dysplasia (**c**) and Arnold–Chiari I + II syndrome (**h**). In the latter, the PONC-derived clivus (blue in **g**) of the otherwise mesodermal (mes) cranial base, which serves as the attachment point (attp) for pharynx constrictor muscles (constr. phar. N.IX/X.) in front of the medulla (med),

fails to form and is replaced by fragile connective tissue. Right: conversely, PONC-derived connective tissues of pharynx constrictor muscles (const. phar. N.IX/X in **g**) that are connected to cervical vertebrae (**e**) can ectopically become bone (**f**), leading to neck immobility in Klippel–Feil syndrome (**f**) or ectopic, ‘omo-vertebral’ bones inside trapezius territory (stippled line in **d**) in patients with Sprengel’s deformity, a frequent facet of Klippel–Feil syndrome^{19,20}. Note also the concomitant loss of the PONC-derived (but not mesodermal) scapular spine in patients with Sprengel’s deformity (**d**).

ossifications of PONC (trapezius) connective tissues around the somitic mesodermal neck vertebrae and an ectopic ossification of the PONC pre-vertebral ligaments of pharyngeal muscles (Fig. 6e, f). Similarly, loss or dysplasia of PONC-derived basicranial (clivus) bone attachments for the internal pharynx and larynx constrictors (Fig. 6g, attp of constr. phar. N.IX) and ensuing widening of the foramen magnum are the primary mechanical cause of the Arnold–Chiari I and II malformation, a serious human congenital malformation associated with swallowing problems and sudden infant (cot) death syndrome (SIDS) (Fig. 6h)²². In this case PONC re-specification from endochondral attachment bone to connective tissue is the likely cause of cryptic basicranial instability and early death. Detailed aetiologies of these congenital syndromes will be discussed elsewhere (T.M. and G.K., unpublished observations). Given the cellular identification of these defects and corresponding anatomical phenotypes of particular transcription factor mutants in mice (Supplementary Fig. S1) the underlying genetic causes can now be investigated in a more focused manner: impaired transcription factor networks acting inside PONC cell populations during development.

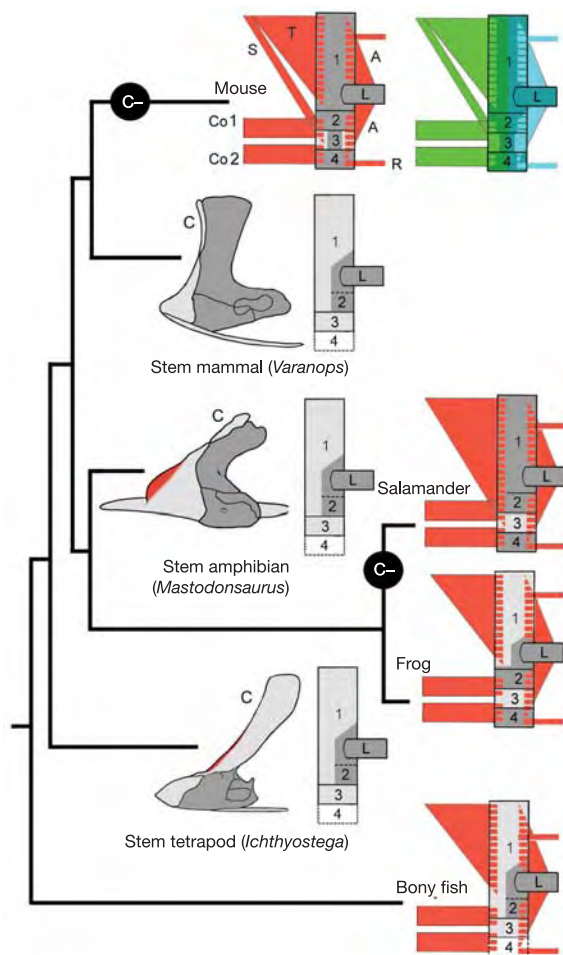


Figure 7 | Muscle scaffolds in fossils. The cleithrum (light grey box 1) has been lost (C-) several times independently in evolution and 'morphs' into scapular spines. In all species with a cleithrum—be they fossil mammalian (*Varanops*³⁸), amphibian (*Mastodonsaurus*³⁷) or tetrapod (*Ichthyostega*³⁵; P.E.A., unpublished observation) stem-group representatives or extant frogs³⁶—it is identical in position to the scapular spine of extant tetrapods with respect to the dual trapezius (T) and mesodermal (A) muscle connectivity and its dorsoventral position. Red zones on fossil forms indicate observed trapezius (T) complex attachment regions. Additional independent cleithrum losses among amniotes have been omitted for clarity (see the text). For abbreviations see Fig. 1.

Discussion

Scaffold model, homologies and mechanisms. We have identified the neck and shoulder region as the interface of the neural crest and mesodermal cell populations. We show that boundaries of embryonic cell populations precisely correspond to muscle attachment regions but not to ossification modes. The evolutionary conservation of muscle patterns (red in Figs 1, 7) is therefore likely to be a reflection of conserved cell population boundaries. The latter seem to be far more stable than the signalling pathways that determine their dermal–endochondral ossification as attachment points (grey and white areas in Fig. 1). An alternative hypothesis would have to find multiple independent developmental explanations for such highly constrained muscle patterns. Verification of cell boundary stability and the validity of the 'scaffold model' will have to await further genetic-fate mapping in a wider phylogenetic range of species when this becomes possible. However, our present high-resolution data set for the mouse allows us to refute the widely held competing 'ossification model'^{9,10}. Dermal versus endochondral ossification modes are not safe criteria for identifying cellular origins and homologies of neck and shoulder structures. The rather counter-intuitive 'scaffold model' perceives muscle connectivities as the basic units (because they precisely correspond to cell populations) but considers the bones that everyone can see as mere epiphenomena and subjects of change. This prompts a new heuristic strategy for experimentally establishing neck homologies on the basis of attachment criteria, details of which can be found in Supplementary Methods S2.

The connectivity patterns that we observe with single-cell resolution in the mesodermal occipital and shoulder girdle are stricter than expected^{14,24}. Muscles are directly connected to skeleton of the same axial *HoxD4*⁺ gene identity, without mediation through connective tissue (blue dots in Figs 3e and 4e, f; arrowheads in Fig. 5g). Our findings corroborate the emerging notion^{25,26} that in vertebrates, as previously found in *Drosophila*²⁷, myoblast (muscle) precursors still harbour positional identity inherited from their somitic mesenchymal stem cell precursors, and are not as naive as commonly perceived. Neural crest and mesoderm independently acquire *Hox* gene expression during evolution²⁸. *Hox* genes have been proposed to be responsible for population contiguity in neural crest^{13,29,30}. We note that mouse mutants of *Runx2*^{+/−} (ref. 21), as well as those of interactors or downstream targets of autonomously acting *Hox* genes such as *Pbx1* (ref. 31), *Pax1* (ref. 32) and *Pax 9* (ref. 33), display modular neck/shoulder defects only of PONC but not mesodermal bones and replicate the human syndromes mentioned above (Supplementary Fig. S1 and data not shown). This supports the notion that cell-type-specific modularity and connectivity are controlled by *Hox* genes. In a complementary manner, *Emx2* mouse mutants lack the mesodermal scapular blade but not PONC shoulder derivatives³⁴. Whereas *HoxD4*⁺ somites also provide muscle cells to the trapezius and other branchial neck muscles, these myoblasts seem to be subjugated to neural-crest-derived muscle connective tissue: they do not attach directly onto skeletal regions as mesodermal muscles do. The challenge will be to disentangle the molecular causes for such patterning dominance of neural crest over mesoderm in areas of spatial overlap, where neural crest and somitic *Hox* codes 'collide'.

Fossil fates: chasing the cleithrum's ghost. The conservation of the neck muscle scaffold among jawed vertebrates and its precise correspondence to cell population boundaries provides refined single-cell criteria for tracing skeletal fate changes of a more fundamental nature. This permits us to determine the whereabouts of elements such as the elusive cleithrum, the centralmost shoulder bone of all bony fish (osteichthyan) ancestors, which is absent from all extant land living vertebrates (tetrapods)³⁵ except frogs³⁶. The cleithrum is uniquely defined by its position and connectivity. In extant bony fish and frogs it serves as the sole attachment region for the trapezius/cucullaris muscles anteriorly^{5,36} and for fin, limb

and trunk muscles at its posterior margin (light grey in box 1 in Figs 1 and 7)^{35,36}. Two main historical hypotheses have been proposed to explain its absence from most living tetrapods: first, the cleithrum was lost in common tetrapod (stem-group) ancestors and re-acquired only in frogs; second, it was independently lost from the lineages leading to living salamanders, diapsids, turtles and mammals. These two hypotheses have divergent implications for understanding macro-evolutionary trends in neck skeletogenesis. Comparative anatomy of extant species cannot distinguish between these two hypotheses (Figs 1 and 7; compare salamander and mammal). Palaeontology, in contrast, has unearthed a rich data set of fossil stem-group shoulder morphologies that can resolve the timing and polarity of these changes. Using these attachment criteria as a guide we re-examine representative examples of stem tetrapods (*Ichthyostega*)³⁵, stem amphibians (*Mastodonsaurus*)³⁷, stem amniotes (*Seymouria* and *Diadectes*, not shown) and stem-group mammals (*Varanops*)³⁸ (Fig. 7). All stem-group tetrapods discovered so far possess a cleithrum (C in Fig. 7) covering the entire antero-dorsal margin of the shoulder girdle. Indeed, trapezius/cucullaris muscle scars can be found on the cleithra of *Jacobsonia*³⁹, *Pederpes* (J. A. Clack and S. M. Finney, personal communication)⁴⁰ and *Ichthyostega* (P.E.A., unpublished observation) (red areas on Fig. 7), which demonstrates a cleithrum muscle connectivity pattern primitively retained from the fish condition. We also find a cleithrum in stem reptiles such as *Araucoscelis*⁴¹ and stem turtles *Proganochelys*⁴² and *Kayentachelys*⁴³. Thus, the cleithrum has been lost at least four times independently: in salamanders, mammals, turtles and diapsids (C— in Fig. 7, and data not shown). Meanwhile, the muscle connectivities embracing the cleithrum have remained unchanged from the fish condition: exactly like the ancestral cleithrum, the mammalian scapular spine is positioned between the conserved trapezius and limb/trunk muscle attachment systems as a common abutment of the two. Based on identical cell-population-mediated connectivity and position, our genetic labelling proposes to identify the endochondral scapular spine as the ‘cell population ghost’ of the cleithrum. It remains to illuminate the molecular causes for this unidirectional trend to dismantle the dermal shoulder girdle, replace it by endochondral skeleton or lose it altogether, which seems to continue in mammals and amphibians and also extends to other bones such as the clavicle^{16,36}. In the framework of the highly constrained neck muscle scaffolds we find no evidence for histogenetic reversals; that is, that endochondral bones of ancestors turned into dermal bones of descendants during the course of evolution. We speculate that a common, as yet unknown, genomic *cis*-regulatory architecture governing neck ossifications in tetrapod ancestors might have predisposed different descending tetrapod lineages to similar parallel trends.

The present study identifies the embryonic cell populations involved in neck patterning: PONC and somitic mesoderm. These mesenchymal stem cell populations are subject to considerable muscle patterning constraints while retaining pathological and evolutionary flexibility in their osteogenic differentiation. The molecular basis of such constraints and flexibilities and their integration into single cells remains to be discovered. Ultimately, this ‘protean’ flexibility of mesenchymal stem cells to ‘morph’ into cartilage, bone and connective tissue will have to be explained in the language of evolving gene-regulatory circuitry. This genetic circuitry will have to be placed into future reconstructions of phylogenetic trees because it was causative for the diverse neck morphologies that we observe. We expect that traces of other major evolutionary transformations and novelties will become detectable on a single-cell level once comparative genetic lineage mapping becomes possible.

METHODS

We generated two independent transgenic mouse lines in which all PONC cells are permanently labelled by means of recombinase-activated marker cassettes. *Wnt1* is expressed in all pre-migratory PONC cell precursors¹⁷ and *Sox10* is

expressed strongly in the entire post-migratory PONC population during early embryonic development and not at all in mesoderm⁴⁴. We therefore labelled pre-migratory PONC with a *Wnt-1-Cre* transgene¹⁷ and post-migratory PONC with a *Sox-10-Cre* bacterial artificial chromosome transgene, introduced into founders by pronuclear injection. Transgenic founders were crossed to Cre-conditional R26-LacZ and R26-eGFP reporter lines^{45,46} (Fig. 2a, left). This allowed us to reveal for the first time the entire LacZ⁺ and GFP⁺ PONC population of the head, neck and shoulder region with single-cell resolution (green in Fig. 2b). Complementary to these lines, we generated a third transgenic line in which all post-occipital paraxial (neck and trunk) mesoderm was permanently labelled, carrying a novel *HoxD4-CREM-LacZ* construct (Fig. 2a, right). Landmark studies on the correspondence between *Hox* gene expression and axial somitic identity had predicted *Hox-4* paralogs to define the non-occipital/occipital boundary^{18,47}. These genetic and embryological studies²⁴ had shown that the occipital plane in amniotes runs exactly through somite 5, which expresses *HoxD4*. *HoxD4* is not expressed in neural crest and its mesodermal expression is regulated by a well-characterized genomic fragment⁴⁸. We cloned this fragment into a novel lineage labelling construct for *HoxD4*⁺ somitic mesenchymal stem cells and all of their progeny from somite 5 backwards (blue in Fig. 2b). In brief, LacZ marker activation was made conditional on a *HoxD4*-enhancer-controlled self-excision of Cre recombinase. This new strategy obviates traditional problems of differential position-effect variegation (PEV in Fig. 2a, right) of transgenic Cre drivers and reporters. It also avoids deleterious effects of high Cre-recombinase levels in tissues⁴⁹. Technical details about constructs, controls and analysis are provided in Supplementary Methods S1.

Received 13 September 2004; accepted 20 May 2005.

1. Janvier, P. *Early Vertebrates* (Oxford Science Publications, Oxford, 1996).
2. Johanson, Z. Placoderm branchial and hypobranchial muscles and origins in jawed vertebrates. *J. Vert. Paleontol.* **23**, 735–749 (2003).
3. Motta, P. J. & Wilga, C. D. *Environmental Biology of Fishes* Vol. 60, 131–156 (Kluwer Academic, Dordrecht, 2001).
4. LeDouarin, N. & Kalcheim, C. *The Neural Crest* 2nd edn (Cambridge Univ. Press, Cambridge, 1999).
5. Edgeworth, F. H. *The Cranial Muscles of Vertebrates* (Cambridge Univ. Press, Cambridge, 1935).
6. Shubin, N., Tabin, C. & Carroll, S. Fossils, genes and the evolution of animal limbs. *Nature* **388**, 639–648 (1997).
7. Lumsden, A., Sprawson, N. & Graham, A. Segmental origin and migration of neural crest cells in the hindbrain region of the chick embryo. *Development* **113**, 1281–1291 (1991).
8. Saunders, J. W. J. The proximo-distal sequence of origin of the parts of the chick wing and the role of the ectoderm. *J. Exp. Zool.* **108**, 363–403 (1948).
9. Smith, M. M. & Hall, B. K. Development and evolutionary origins of vertebrate skeletogenic and odontogenic tissues. *Biol. Rev.* **65**, 277–373 (1990).
10. Smith, M. M. & Hall, B. K. In *Evolutionary Biology* Vol. 27 (eds Hecht, M. K., MacIntyre, R. J. & Clegg, M. T.) 387–448 (Plenum, New York, 1993).
11. Couly, G. F., Coltey, P. M. & LeDouarin, N. M. The triple origin of skull in higher vertebrates: a study in quail–chick chimeras. *Development* **114**, 1–15 (1993).
12. Jiang, X., Iseki, S., Maxson, R. E., Sucov, H. M. & Morriss-Kay, G. M. Tissue origins and interactions in the mammalian skull vault. *Dev. Biol.* **241**, 106–116 (2002).
13. Koentges, G. & Lumsden, A. G. S. Rhombencephalic neural crest segmentation is preserved throughout craniofacial ontogeny. *Development* **122**, 3229–3242 (1996).
14. Huang, R. et al. Contribution of single somites to the skeleton and muscles of the occipital and cervical regions in avian embryos. *Anat. Embryol.* **202**, 375–383 (2000).
15. Huang, L. F. et al. Mouse clavicular development: analysis of wild-type and cleidocranial dysplasia mutant mice. *Dev. Dyn.* **210**, 33–40 (1997).
16. Hall, B. K. Development of the clavicles in birds and mammals. *J. Exp. Zool.* **289**, 153–161 (2001).
17. Danielian, P. S., Muccino, D., Rowitch, D. H., Michael, S. K. & McMahon, A. P. Modification of gene activity in mouse embryos *in utero* by a tamoxifen-inducible form of Cre recombinase. *Curr. Biol.* **8**, 1323–1326 (1998).
18. Burke, A. C., Nelson, C. E., Morgan, B. A. & Tabin, C. *Hox* genes and the evolution of vertebrate axial morphology. *Development* **121**, 333–346 (1995).
19. Clarke, R. A., Catalan, G., Diwan, A. D. & Kearsley, J. H. Heterogeneity in Klippel-Feil syndrome: a new classification. *Pediatr. Radiol.* **28**, 967–974 (1998).
20. Horwitz, A. E. Congenital elevation of the scapula—Sprengel's deformity. *Am. J. Orthop. Surg.* **6**, 260–311 (1908).
21. Otto, F. et al. *Cbfa1*, a candidate gene for cleidocranial dysplasia syndrome, is essential for osteoblast differentiation and bone development. *Cell* **89**, 765–771 (1997).
22. Graham, D. I. & Lantos, P. L. *Greenfield's Neuropathology* 7th edn (Oxford Univ. Press, London, 2002).

23. Kjaer, I. & Niebuhr, E. Studies of the cranial base in 23 patients with cri-du-chat syndrome suggest a cranial developmental field involved in the condition. *Am. J. Med. Genet.* **82**, 6–14 (1999).
24. Huang, R., Zhi, Q., Patel, K., Wilting, J. & Christ, B. Dual origin and segmental organisation of the avian scapula. *Development* **127**, 3789–3794 (2000).
25. Alvares, L. E. *et al.* Intrinsic, Hox-dependent cues determine the fate of skeletal muscle precursors. *Dev. Cell* **5**, 379–390 (2003).
26. Schweitzer, R. *et al.* Analysis of the tendon cell fate using Scleraxis, a specific marker for tendons and ligaments. *Development* **128**, 3855–3866 (2001).
27. Baylies, M. K. *et al.* Myogenesis: a view from *Drosophila*. *Cell* **93**, 921–927 (1998).
28. Takio, Y. *et al.* Lamprey Hox genes and the evolution of jaws. *Nature* **429**, 262–263 (2004).
29. Barrow, J. R. & Capecchi, M. R. Compensatory defects associated with mutations in *Hoxa1* restore normal palatogenesis to *Hoxa2* mutants. *Development* **126**, 5011–5026 (1999).
30. Smith, A. *et al.* The EphA4 and EphB1 receptor tyrosine kinases and ephrin-B2 ligand regulate targeted migration of branchial neural crest cells. *Curr. Biol.* **7**, 561–570 (1997).
31. Salleri, L. *et al.* Requirement for Pbx1 in skeletal patterning and programming chondrocyte proliferation and differentiation. *Development* **128**, 3543–3557 (2001).
32. Dietrich, S. & Gruss, P. Undulated phenotypes suggest a role of Pax-1 for the development of vertebral and extravertebral structures. *Dev. Biol.* **167**, 529–548 (1995).
33. Peters, H. *et al.* Pax1 and Pax9 synergistically regulate vertebral column development. *Development* **126**, 5399–5408 (1999).
34. Prols, F. *et al.* The role of Emx2 during scapula formation. *Dev. Biol.* **275**, 315–324 (2004).
35. Jarvik, E. *Basic Structure and Evolution of Vertebrates* Vol. 1 (Academic, London, 1980).
36. Shearman, R. M. Growth of the pectoral girdle of the Leopard Frog *Rana pipiens* (Anura: Ranidae). *J. Morphol.* **264**, 94–104 (2005).
37. Schoch, R. R. Comparative osteology of *Mastodonsaurus giganteus* (Jaeger, 1828) from the Middle Triassic (Lettenkeuper: Longobardian) of Germany (Baden-Württemberg, Bayern, Thüringen). *Stuttg. Beitr. Naturk. B* **278**, 1–175 (1999).
38. Sumida, S. S. in *Amniote Origins* (eds Sumida, S. S. & Martin, K. L. M.) 353–398 (Academic, San Diego, 1997).
39. Lebedev, O. A. in *The Second Gross Symposium 'Advances in Palaeoichthyology'* (ed. Luksevics, E.) 79–98 (*Acta Universitatis Latviensis* 679, 2005).
40. Clack, J. A. & Finney, S. M. *Pederpes finneyae*, an articulated tetrapod from the Tournaisian of Western Scotland. *J. Syst. Palaeont.* **2**, 311–346 (2005).
41. Reisz, R. R., Berman, D. S. & Scott, D. The anatomy and relationships of the Lower Permian reptile *Araucoscelis*. *J. Vertebr. Paleontol.* **4**, 57–67 (1984).
42. Jaekel, O. Die Wirbeltierfunde aus dem Keuper von Halberstadt. *Paläont. Z.* **2**, 88–214 (1915–16).
43. Joyce, W. The presence of cleithra in the primitive turtle *Kayentachelys aprix*. *J. Vert. Paleontol.* **23** (suppl.), 66A (2003).
44. Ferguson, C. A. & Graham, A. Redefining the head–trunk interface for the neural crest. *Dev. Biol.* **269**, 70–80 (2004).
45. Soriano, P. Generalized lacZ expression with the ROSA26 Cre reporter strain. *Nature Genet.* **21**, 70–71 (1999).
46. Mao, X. *et al.* Activation of EGFP expression by Cre-mediated excision in a new ROSA26 reporter mouse strain. *Blood* **97**, 324–326 (2001).
47. Condie, B. G. & Capecchi, M. R. Mice with targeted disruptions in the paralogous genes *Hoxa-3* and *Hoxd-3* reveal synergistic interactions. *Nature* **370**, 304–307 (1994).
48. Zhang, F. *et al.* Elements both 5' and 3' to the murine *Hoxd4* gene establish anterior borders of expression in mesoderm and neuroectoderm. *Mech. Dev.* **67**, 49–58 (1997).
49. Loonstra, A. *et al.* Growth inhibition and DNA damage induced by Cre recombinase in mammalian cells. *Proc. Natl Acad. Sci. USA* **98**, 9209–9214 (2001).
50. Winterbottom, R. A descriptive synonymy of the striated muscles of the teleostei. *Proc. Acad. Nat. Sci. Phila.* **125**, 225–317 (1974).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Lumsden for help with a complex manuscript; P. Soriano and S. Orkin for providing Cre reporters; and A. West, G. Felsenfeld and J. Green for advice on insulators and plasmids. This work was funded by the BBSRC (G.K., P.E.A.), the Wellcome Trust (G.K., W.D.R.), the MRC UK (W.D.R.), the Swedish Research Council (P.E.A.), the NIH (A.P.M.) and WIBR-UCL (G.K.). G.K. and T.M. were long-term postdoctoral fellows of HFSPO. G.K. thanks S. Moncada for support in establishing a new laboratory.

Author Contributions T.M. and P.E.A. contributed equally to this work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.K. (g.koentges@ucl.ac.uk).

ARTICLES

Serum retinol binding protein 4 contributes to insulin resistance in obesity and type 2 diabetes

Qin Yang^{1*}, Timothy E. Graham^{1*}, Nimesh Mody¹, Frederic Preitner¹, Odile D. Peroni¹, Janice M. Zabolotny¹, Ko Kotani¹, Loredana Quadro^{2†} & Barbara B. Kahn¹

In obesity and type 2 diabetes, expression of the GLUT4 glucose transporter is decreased selectively in adipocytes. Adipose-specific *Glut4* (also known as *Slc2a4*) knockout (adipose-*Glut4*^{-/-}) mice show insulin resistance secondarily in muscle and liver. Here we show, using DNA arrays, that expression of retinol binding protein-4 (RBP4) is elevated in adipose tissue of adipose-*Glut4*^{-/-} mice. We show that serum RBP4 levels are elevated in insulin-resistant mice and humans with obesity and type 2 diabetes. RBP4 levels are normalized by rosiglitazone, an insulin-sensitizing drug. Transgenic overexpression of human RBP4 or injection of recombinant RBP4 in normal mice causes insulin resistance. Conversely, genetic deletion of *Rbp4* enhances insulin sensitivity. Fenretinide, a synthetic retinoid that increases urinary excretion of RBP4, normalizes serum RBP4 levels and improves insulin resistance and glucose intolerance in mice with obesity induced by a high-fat diet. Increasing serum RBP4 induces hepatic expression of the gluconeogenic enzyme phosphoenolpyruvate carboxykinase (PEPCK) and impairs insulin signalling in muscle. Thus, RBP4 is an adipocyte-derived 'signal' that may contribute to the pathogenesis of type 2 diabetes. Lowering RBP4 could be a new strategy for treating type 2 diabetes.

A major cause of type 2 diabetes is impaired insulin action in adipose tissue, skeletal muscle and liver. Overt hyperglycaemia develops when increased insulin secretion no longer compensates for insulin resistance. Even without diabetes, insulin resistance is a major risk factor for cardiovascular disease and early mortality¹. Resistance to insulin-stimulated glucose transport in adipose tissue and skeletal muscle is one of the earliest defects detected in insulin-resistant states². Transmembrane transport of glucose by GLUT4 is the rate-limiting step for glucose use by muscle and adipose tissue². With the development of insulin resistance, GLUT4 expression is downregulated selectively in adipose tissue but not in skeletal muscle^{2,3}. Downregulation of GLUT4 expression in adipose tissue is an almost universal feature of insulin-resistant states, including obesity, type 2 diabetes and the metabolic syndrome².

Skeletal muscle is regarded as the principal site for insulin-stimulated glucose uptake, whereas adipose tissue takes up much less glucose under normal physiological conditions³. To determine how GLUT4 expression in adipose tissue influences systemic insulin action, we generated transgenic mice with adipose-specific overexpression of human GLUT4 (adipose-GLUT4-Tg mice)⁴ or adipose-specific reduction of GLUT4 using Cre/loxP gene targeting (adipose-*Glut4*^{-/-} mice)⁵. These mice show opposite phenotypes with regards to insulin sensitivity and glucose homeostasis. Adipose-GLUT4-Tg mice have enhanced glucose tolerance and insulin sensitivity⁴. Relatively increased insulin-sensitivity persists even in the diabetic state induced by pancreatic β -cell destruction⁶, and GLUT4 overexpression in adipocytes of mice lacking GLUT4 in muscle reverses their diabetes⁷. Thus, increasing GLUT4 expression

selectively in adipocytes protects against whole-body insulin resistance. In contrast, mice with markedly reduced GLUT4 expression in adipose tissue, but normal GLUT4 expression in muscle, are insulin-resistant and have an increased risk of overt diabetes⁵. Adipose-specific deletion of *Glut4* leads to secondary defects in insulin action in muscle and liver⁵. However, insulin action in muscle of adipose-*Glut4*^{-/-} mice *ex vivo* is normal⁵, indicating that a circulating factor(s) causes insulin resistance in these mice. We sought to identify such a factor(s) released from adipocytes.

Although adipose tissue is traditionally regarded as an inert energy-storage depot, it is also an active endocrine organ⁸. Adipocyte-secreted molecules can either enhance (for example, leptin and adiponectin) or impair (for example, fatty acids, tumour-necrosis factor- α (TNF- α) and resistin) insulin action. Serum levels of most adipocyte-secreted molecules known to influence insulin action are normal in adipose-*Glut4*^{-/-} mice (ref. 5 and unpublished data). Thus, insulin resistance in muscle and liver of adipose-*Glut4*^{-/-} mice is probably caused by altered secretion of an unknown adipocyte-derived molecule(s).

We therefore undertook global gene expression analysis to identify genes for which expression is altered in adipose tissue harbouring a primary genetic alteration in GLUT4 expression. Here we report that serum retinol binding protein (UniGene RBP4) is upregulated in adipose tissue of adipose-*Glut4*^{-/-} mice, and that serum RBP4 levels are elevated in several insulin-resistant states in mice and humans. RBP4 is the only specific transport protein for retinol (vitamin A) in the circulation, and to date, its only known function is to deliver retinol to tissues⁹. We find that elevation of serum RBP4

¹Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA. ²Institute of Cancer Research, Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA. [†]Present address: Department of Food Science, Rutgers-The State University of New Jersey, New Brunswick, New Jersey 08901, USA.

*These authors contributed equally to this work.

causes systemic insulin resistance, and that reduction of serum RBP4 improves insulin action. These results point to a novel mechanism underlying the inter-tissue communication that has been shown to have an important role in the pathogenesis of type 2 diabetes^{10,11}.

DNA array reveals adipose-RBP4 regulation

We performed DNA array analyses on epididymal adipose tissue RNA from adipose-*Glut4*^{-/-} (ref. 5) and adipose-*GLUT4*-Tg (ref. 4) mice using Affymetrix mouse MG-U74Av.2 oligonucleotide arrays containing ~12,000 genes. We identified genes that were reciprocally

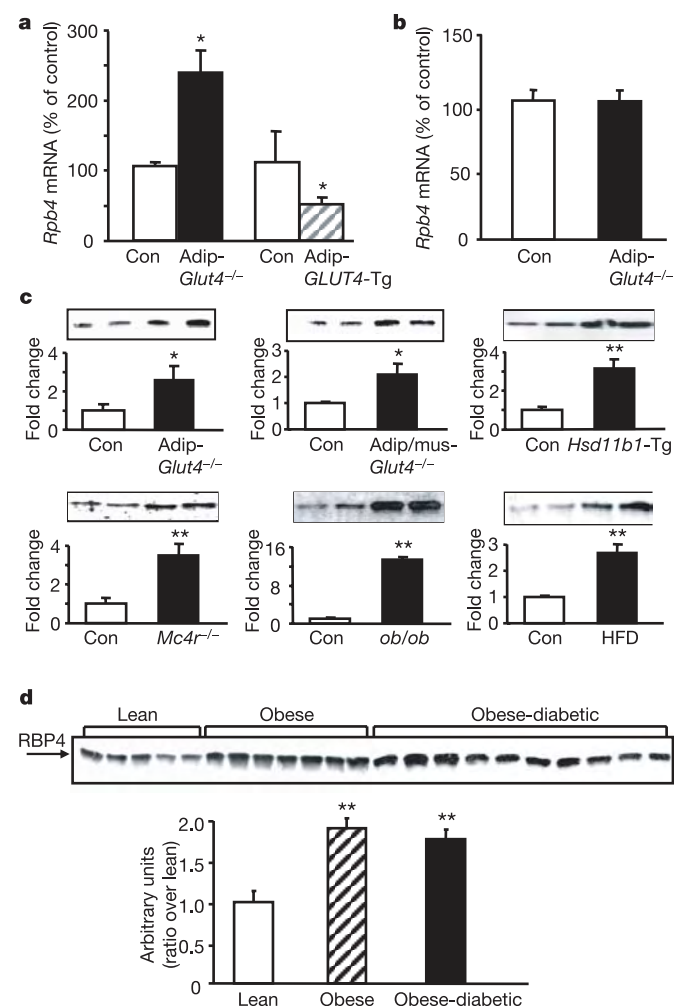


Figure 1 | Elevation of RBP4 in insulin-resistant mouse models. **a**, *Rbp4* mRNA levels in adipose tissue of adipose-*Glut4*^{-/-} and adipose-*GLUT4*-Tg mice and their controls (Con) at 14 weeks of age ($n = 3-4$ per genotype). Asterisk denotes $P < 0.05$ versus controls. mRNA was measured by one-step Taqman real-time PCR. **b**, *Rbp4* mRNA levels in liver of adipose-*Glut4*^{-/-} and control mice at 12–14 weeks ($n = 7-8$ per genotype). **c**, Serum levels of RBP4 in insulin-resistant mouse models. RBP4 was detected by western blotting⁹, and graphs show densitometric quantification. Mice were 8–16 weeks of age ($n = 3-5$ mice per condition or genotype). Mice fed on a high-fat diet (HFD) for 12 weeks obtained 55% of their calories from fat. Body weight (g) for chow diet (mean $\pm$ s.e.m.), 32.0 ± 0.8 ; HFD, 37.5 ± 0.3 . Plasma insulin levels (ng ml⁻¹) for chow diet: 1.4 ± 0.1 ; HFD: 8.3 ± 0.6 . Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ versus controls. Adip/mus-*Glut4*^{-/-}, *Glut4* knockout in both adipose tissue and muscle¹³; *Hsd11b1*-Tg, adipose overexpression of hydroxysteroid 11- β dehydrogenase-1; *Mc4r*^{-/-}, melanocortin 4 receptor knockout. **d**, Serum RBP4 in humans. Each lane contains serum from a different person. Graph (lower panel) shows densitometric quantification of bands. Two asterisks, $P < 0.01$ versus lean non-diabetic subject. Data in **a–d** are presented as mean $\pm$ s.e.m.

regulated by these genetic manipulations and that encoded secreted proteins. Five messenger RNAs encoding known secreted proteins were inversely regulated (an approximately twofold change) in adipose tissue of adipose-*Glut4*^{-/-} and adipose-*GLUT4*-Tg mice. One was serum retinol binding protein (RBP4). Real-time quantitative polymerase chain reaction (PCR) showed that *Rbp4* mRNA was increased 2.3-fold in adipose tissue of adipose-*Glut4*^{-/-} mice, and was reduced by 54% in adipose tissue of adipose-*GLUT4*-Tg mice (Fig. 1a). *Rbp4* is normally expressed at very high levels in liver¹², but hepatic *Rbp4* expression was unaltered in adipose-*Glut4*^{-/-} mice (Fig. 1b).

Elevated serum RBP4 in insulin resistance

Consistent with the mRNA changes in adipose tissue, serum RBP4 in adipose-*Glut4*^{-/-} mice was elevated 2.5-fold compared with control mice (Fig. 1c). Serum RBP4 levels were also increased in five other insulin-resistant mouse models (Fig. 1c): mice with reduction of *Glut4* in both adipose tissue and muscle¹³ (2-fold increase); mice overexpressing 11- β hydroxysteroid dehydrogenase-1 in adipose tissue (adipose-*Hsd11b1*-Tg)¹⁴ (3-fold increase), melanocortin 4 receptor (*Mc4r*) knockout mice¹⁵ (3.5-fold increase); mice on a high-fat (55% fat) diet (2.8-fold increase) and ob/ob mice (13-fold increase) (Fig. 1c). Plasma RBP4 correlated with insulin levels in the fed state in obese mice on a high-fat diet ($r^2 = 0.83$, $P < 0.01$; see Supplementary Fig. 1).

We measured plasma RBP4 levels in obese humans with or without type 2 diabetes. Obese-diabetic subjects were older than obese-nondiabetic or lean control subjects¹⁶ (see Supplementary Table 1).

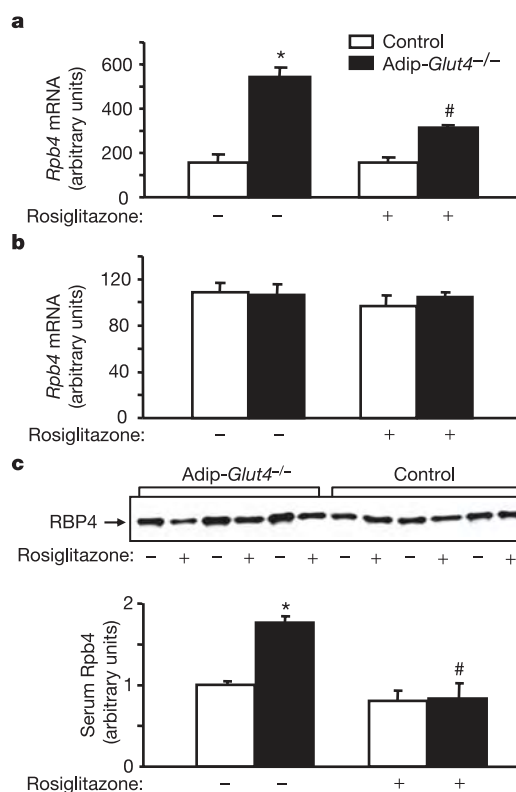


Figure 2 | Rosiglitazone treatment decreases RBP4 levels in adipose-*Glut4*^{-/-} mice. **a**, **b**, *Rbp4* mRNA in adipose tissue (**a**) or liver (**b**) of adipose-*Glut4*^{-/-} and control female mice age 27 weeks treated with (+) or without (-) rosiglitazone ($n = 3-6$ per condition). **c**, Serum RBP4 in adipose-*Glut4*^{-/-} and control mice before (-) and after (+) 3 weeks of rosiglitazone treatment ($n = 4$ per condition). Asterisk, $P < 0.05$ versus control (-); hash symbol, $P < 0.05$ versus adipose-*Glut4*^{-/-} (-). Data in **a–c** are presented as mean $\pm$ s.e.m.

Obese and obese-diabetic subjects had higher body mass index (BMI) and fasting insulin levels and lower glucose disposal rates (measured by euglycemic-hyperinsulinemic clamp studies) compared with lean controls. Fasting glucose and haemoglobin A_{1c} (HbA_{1c}, which reflects average glucose levels in blood over 2–3 months) were elevated in obese-diabetic subjects compared with non-diabetic subjects (see Supplementary Table 1). Serum RBP4 levels were increased ~1.9-fold in obese-nondiabetic and obese-diabetic subjects compared with lean controls (Fig. 1d). There was no difference in the magnitude of serum RBP4 elevation between the obese and obese-diabetic groups, suggesting that obesity and insulin resistance, but not hyperglycaemia, are associated with elevated serum RBP4 in humans.

Rosiglitazone reverses RBP4 elevation

The anti-diabetic agent rosiglitazone is a peroxisome proliferator activated receptor- γ (PPAR γ) agonist that improves insulin sensitivity¹⁷. Treatment of adipose-*Glut4*^{-/-} mice with rosiglitazone for three weeks completely reversed their insulin resistance and glucose intolerance¹⁸. In parallel, rosiglitazone treatment reduced the elevated *Rbp4* mRNA levels in adipose tissue (Fig. 2a) but not in liver (Fig. 2b) of adipose-*Glut4*^{-/-} mice. Rosiglitazone treatment also completely normalized the elevated serum RBP4 levels in adipose-*Glut4*^{-/-} mice (Fig. 2c). The dramatic effect of this insulin-sensitizing anti-diabetic agent on serum RBP4 levels raises the possibility that elevation of RBP4 might play a causative role in insulin resistance and type 2 diabetes.

Increased RBP4 causes insulin resistance

Transgenic mice expressing human RBP4 driven by the mouse muscle

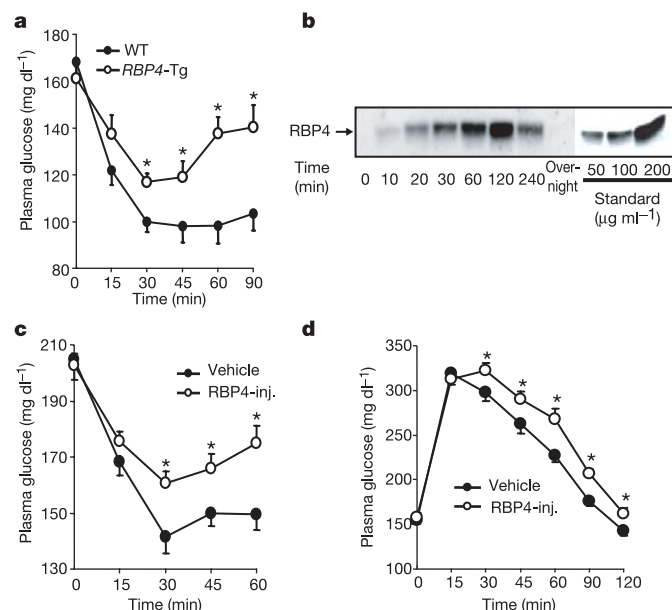


Figure 3 | Elevated serum RBP4 causes insulin resistance. **a**, Insulin tolerance tests in male mice expressing human RBP4 in muscle (*RBP4*-Tg) at 12 weeks of age ($n = 7$ per genotype). Insulin (0.9 U kg^{-1} body weight) was injected intraperitoneally (i.p.) 4 h after food removal. Plasma glucose was measured at the indicated times. Asterisk, $P < 0.01$ versus wild type (WT). **b**, Western blot of serum from an FVB mouse injected with 0.5 mg recombinant human RBP4 i.p. Purified RBP4 was diluted to the indicated concentrations as the standard. **c**, **d**, Insulin (**c**) and glucose (**d**) tolerance tests in 8-week-old mice injected chronically with purified RBP4. Control mice were injected with dialysate (vehicle). Glucose (1 g kg^{-1} i.p.) and insulin (0.9 U kg^{-1} i.p.) tolerance tests were performed after 9 or 19 days of injections, respectively ($n = 8$ control and 12 RBP4-injected mice). Asterisk denotes a significant difference at $P < 0.01$. Data are presented as mean $\pm$ s.e.m.

creatinine kinase (*Mck*) promoter (*RBP4*-Tg) have an ~3-fold increase in serum RBP4 levels compared with non-transgenic mice¹⁹, similar to the elevation observed in serum of adipose-*Glut4*^{-/-} mice (see Supplementary Fig. 2). *RBP4*-Tg mice develop normally. Growth curves are similar to wild-type mice until at least 16 weeks of age (not shown). Insulin levels are higher in fed *RBP4*-Tg mice compared with wild-type mice, indicating insulin resistance (see Supplementary Table 2), which is also evident in insulin tolerance tests (Fig. 3a). There are no differences in glucose, free fatty acid, leptin, adiponectin or resistin levels in fed *RBP4*-Tg mice compared with controls (see Supplementary Table 2).

To circumvent the possibility of developmental or compensatory effects of RBP4 overexpression in transgenic mice, we purified recombinant human RBP4 and injected it into normal, adult FVB mice. RBP4 is a 21-kDa protein that is easily filtered through the renal glomerular membrane. In the circulation, RBP4 binds transthyretin (TTR) to form an 80-kDa protein complex that prevents renal clearance of RBP4 (ref. 20). To determine the pharmacodynamics of exogenous RBP4, we injected 0.5 mg recombinant human RBP4 intraperitoneally and measured serum RBP4 levels using an anti-human RBP4 antibody (Fig. 3b). RBP4 could be detected 10 min

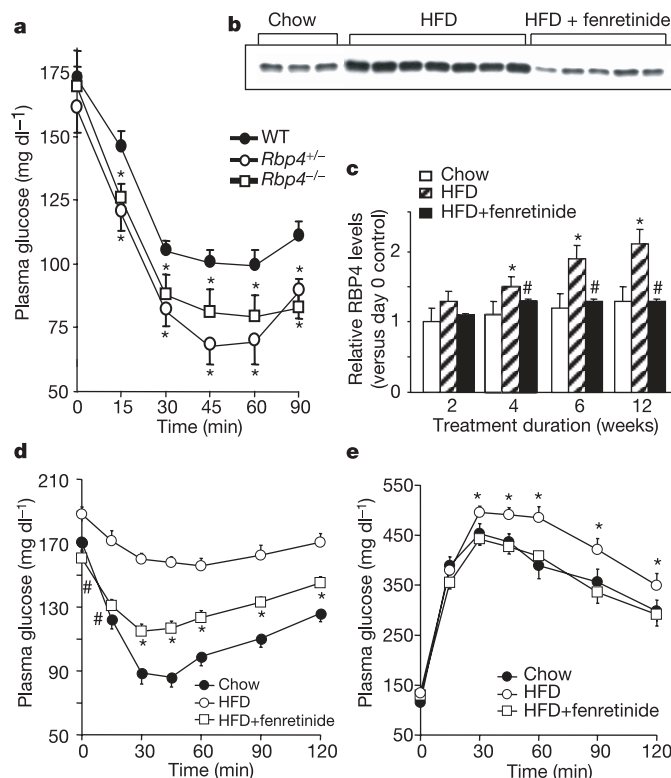


Figure 4 | Lowering serum RBP4 levels improves insulin action. **a**, Insulin tolerance tests in 10-week-old male *Rbp4*^{+/-} and *Rbp4*^{-/-} mice. Insulin (0.75 U kg^{-1}) was injected i.p. ($n = 7$ –9 per genotype). Asterisk, $P < 0.01$ versus wild type at each time point. **b**, Representative immunoblot of serum RBP4 levels in mice fed for 6 weeks with chow diet, high-fat diet (HFD) or HFD with 0.04% fenretinide. **c**, Densitometric quantification of immunoblots of serum RBP4 levels in mice fed chow, HFD or HFD + fenretinide for the indicated times ($n = 8$ –12 per condition). Asterisk, $P < 0.05$ versus chow diet; hash symbol, $P < 0.05$ versus HFD. **d**, Insulin tolerance test (1.1 U kg^{-1} insulin i.p.) in mice fed chow, HFD or HFD + fenretinide for 15 weeks ($n = 6$ –10 mice per condition). Asterisk, $P < 0.05$ versus HFD and chow; hash symbol, $P < 0.05$ versus HFD only. **e**, Glucose tolerance test (2 g kg^{-1} glucose i.p.) in mice fed chow, HFD or HFD + fenretinide for 16 weeks ($n = 8$ –12 mice per condition). Asterisk denotes $P < 0.01$ versus chow and HFD + fenretinide. There was no difference between HFD + fenretinide and chow groups. Data are presented as mean $\pm$ s.e.m.

after injection, and levels peaked at 120 min. At 240 min, the level was 25% of the peak level, indicating rapid clearance. Levels of RBP4 were hardly detectable in serum 16 h after injection (Fig. 3b).

The concentration of endogenous mouse RBP4 in serum is $\sim 30\text{--}40\ \mu\text{g ml}^{-1}$ (ref. 9). To determine whether elevation of RBP4 causes insulin resistance in normal mice, we injected $300\ \mu\text{g}$ purified human RBP4 per mouse per day, as three divided doses ($3\text{--}4\ \mu\text{g g}^{-1}$ body weight) at 8–10-h intervals. This resulted in a daily average serum level of human RBP4 approximately three times higher than endogenous mouse RBP4 (see Supplementary Fig. 2). Control mice were injected with the same volume of dialysate, obtained during the final step of RBP4 purification. RBP4 injection for 9–21 days caused insulin resistance (Fig. 3c) and glucose intolerance (Fig. 3d).

Rbp4 knockout mice ($Rbp4^{-/-}$) are viable and fertile, with normal body weight when maintained on a vitamin A-sufficient diet (see Supplementary Table 2), but have reduced blood retinol levels and impaired visual function early in life⁹. Serum RBP4 is absent in $Rbp4$ homozygous knockout ($Rbp4^{-/-}$) mice and is reduced by 50% in $Rbp4$ heterozygous knockout ($Rbp4^{+/-}$) mice compared with wild-type littermates (see Supplementary Fig. 2). Food intake is normal (not shown). Free fatty acid levels in serum are lower in both $Rbp4^{+/-}$ and $Rbp4^{-/-}$ mice relative to controls (see Supplementary Table 2). Insulin, glucose, leptin, adiponectin and resistin levels in fed mice are normal (see Supplementary Table 2). Both $Rbp4^{+/-}$ and $Rbp4^{-/-}$ mice showed enhanced insulin sensitivity (Fig. 4a). As Rbp4 has been detected in pancreatic α -cells, but

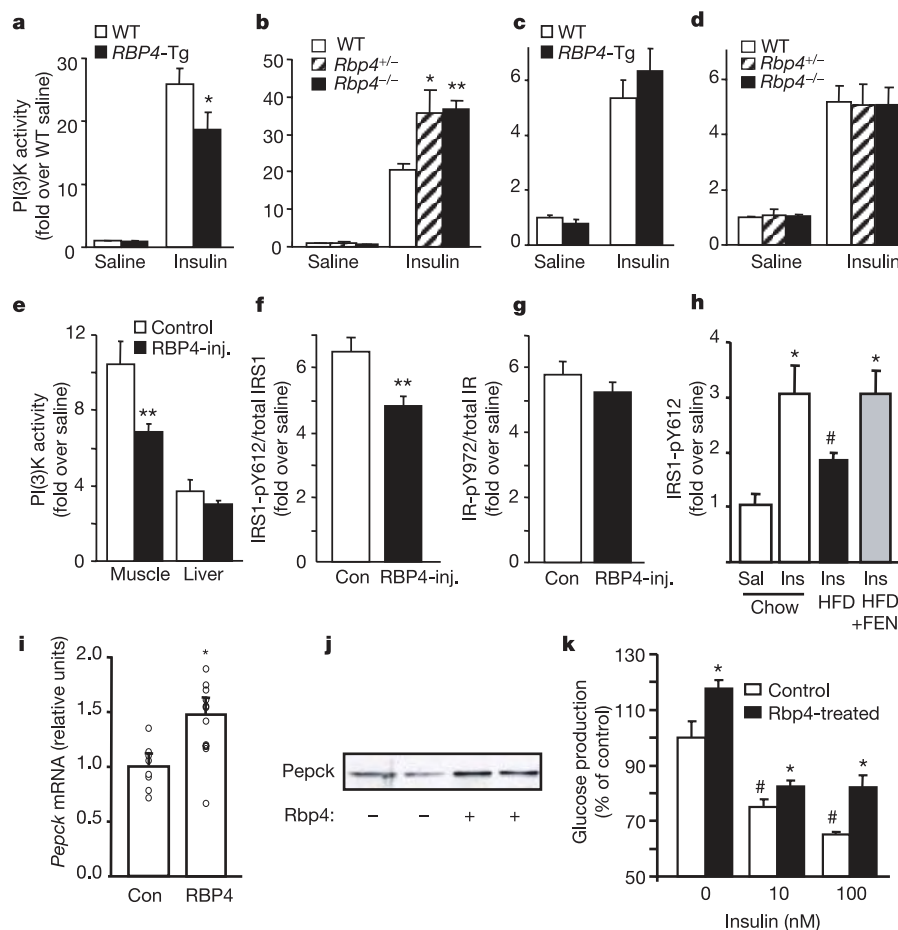


Figure 5 | Effects of RBP4 on insulin signalling, hepatic PEPCK expression and hepatocyte glucose production. **a–d**, PI(3)K activity in muscle (**a**, **b**) or liver (**c**, **d**) of saline or insulin-injected $RBP4$ -Tg mice (**a**, **c**) and $Rbp4^{-/-}$ mice (**b**, **d**) at 16 weeks of age ($n = 4$ saline, $n = 6$ insulin). PI(3)K activity was measured in anti-phosphotyrosine immunoprecipitates. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ versus wild-type (WT) insulin. **e**, Insulin-stimulated PI(3)K activity in muscle and liver of normal FVB mice injected with purified human RBP4 i.p. for 21 days ($n = 4\text{--}9$ per condition). PI(3)K activity in saline-injected mice did not differ between groups. Data are expressed as fold stimulation by insulin over the basal (saline-injected) level. Two asterisks, $P < 0.01$ for RBP4-injected mice versus vehicle control. **f**, **g**, Insulin-stimulated tyrosine phosphorylation of IRS-1 on Y612 (IRS1-pY612) (**f**) and IR (IR-pY972) (**g**) in muscle of normal FVB mice injected with purified human RBP4 ($n = 4\text{--}9$ per condition). Basal (saline-injected) levels of IR and IRS1 phosphorylation did not differ between groups. Data were corrected for the total amount of IRS1 or IR protein in each sample and are expressed as fold stimulation by insulin over the basal (saline-injected) level. Two asterisks, $P < 0.01$ for RBP4-injected mice versus vehicle control. **h**, Insulin-stimulated tyrosine phosphorylation of

IRS-1 in muscle of FVB mice fed a chow diet ($n = 6$), high-fat diet (HFD, $n = 9$) or HFD containing 0.04% (w/w) fenretinide (HFD+FEN, $n = 7$). The basal (saline-injected) level of IRS1 phosphorylation did not differ between groups. Data are expressed as fold stimulation by insulin over the basal (saline-injected) level. Asterisk, $P < 0.05$ versus chow saline; hash symbol, $P < 0.05$ versus chow insulin and HFD+FEN insulin. **i**, *Pepck* mRNA measured by northern blotting in liver of normal FVB mice injected with purified human RBP4 for 21 days. Mice were killed in the overnight fasted state ($n = 8$ for vehicle-treated control mice and 12 for RBP4-treated mice). Asterisk, $P < 0.01$ versus control. **j**, PEPCK protein levels in H4IIE cells treated with purified mouse RBP4 ($100\ \mu\text{g ml}^{-1}$ for 18 h). Western blotting was performed using anti-rat PEPCK antibody²⁸. **k**, Glucose production in H4IIE cells treated for 18 h with vehicle control or purified mouse RBP4 then treated for an additional 3 h with insulin ($n = 3$ per condition). Hash symbol, $P < 0.01$ for the suppressive effect of insulin on glucose production versus cells not treated with insulin; asterisk, $P < 0.05$ for RBP4 treated cells versus control cells at the same insulin concentration. Data are presented as mean $\pm$ s.e.m.

not in β -cells²¹, we measured plasma glucagon levels in *Rbp4*^{-/-} mice. Glucagon levels were normal in the fasting state, and increased normally in response to insulin-induced hypoglycaemia (not shown).

To determine the effect of lowering RBP4 independently of transgenic manipulation, we treated adult FVB mice with fenretinide (4-(*N*-hydroxyphenyl) retinamide), a synthetic retinoid designed for cancer therapy. The bulky side chain (hydroxyphenyl group) disrupts the interaction of RBP4 with TTR, causing renal excretion of RBP4 and resulting in lower serum RBP4 levels²². Oral administration of fenretinide lowered serum RBP4 levels in obese mice on a high-fat diet to the level in lean control mice on a chow diet (Fig. 4b, c). Fenretinide treatment did not affect food intake, body weight or the development of obesity on the high-fat diet (not shown). Mice on a high-fat diet developed marked insulin resistance (Fig. 4d) and glucose intolerance (Fig. 4e). Fenretinide treatment improved insulin sensitivity (Fig. 4d) and normalized glucose tolerance (Fig. 4e). Similar results were obtained in ob/ob mice treated with fenretinide (not shown). Thus, both genetic and pharmacological interventions that decrease serum RBP4 levels improve insulin sensitivity.

RBP4 impairs insulin signalling in muscle

To understand how RBP4 alters insulin sensitivity, we studied insulin signalling in muscle and liver of *Rbp4*^{-/-} mice and mice over-expressing human RBP4 (*RBP4*-Tg mice). Basal phosphoinositide 3-kinase (PI(3)K) activity was similar in all genotypes (Fig. 5a–d). Insulin resulted in a 26-fold increase in PI(3)K activity in muscle of control mice, but its effect was reduced by 30% in *RBP4*-Tg mice (Fig. 5a). Conversely, insulin-stimulated PI(3)K activity was increased by 80% in muscle of both *Rbp4*^{+/-} and *Rbp4*^{-/-} mice compared with control mice (Fig. 5b). However, PI(3)K activity was not altered in the liver of *RBP4*-Tg (Fig. 5c) or *Rbp4*^{-/-} mice (Fig. 5d). Consistent with these observations, RBP4 injection for 21 days in wild-type mice caused a 34% reduction in insulin-stimulated PI(3)K activity in muscle, but no alteration in liver (Fig. 5e). Furthermore, RBP4 treatment resulted in a 24% reduction in insulin-stimulated tyrosine phosphorylation of insulin receptor substrate-1 (IRS1) at tyrosine residue 612 (Fig. 5f), an important docking site for the p85 subunit of PI(3)K²³. However, RBP4 treatment did not alter insulin receptor (IR) tyrosine phosphorylation (Fig. 5g) or the total amount of IRS1 or IR proteins.

Consistent with this, treatment of obese, insulin-resistant mice on a high-fat diet with fenretinide normalized serum RBP4 levels and restored phosphorylation of IRS-1 (on Y612) in muscle to the level in non-obese control mice (Fig. 5h). Total IRS1 levels tended to be higher ($34.5 \pm 0.22\%$, $P = 0.055$) in mice on a high-fat diet treated with fenretinide, compared with both chow-fed mice and mice on a high-fat diet not receiving fenretinide; this might contribute to the increased levels of phosphorylated IRS1. IR expression and phosphorylation were not altered by fenretinide treatment of mice on a high-fat diet (not shown). These data suggest that RBP4 alters insulin sensitivity in part by affecting insulin signalling in muscle through alterations in the amount of tyrosine-phosphorylated IRS-1 and PI(3)K activation. Similar post-receptor defects were observed in the muscle of adipose-*Glut4*^{-/-} mice (ref. 5 and unpublished data), consistent with the notion that elevated serum RBP4 contributes to systemic insulin resistance in this model of type 2 diabetes. However, insulin-stimulated PI(3)K activity was also impaired in liver of adipose-*Glut4*^{-/-} mice, suggesting that RBP4 might not be the only molecule affecting insulin action in these mice.

RBP4 increases hepatic glucose output

In spite of the fact that alterations in plasma RBP4 levels do not affect PI(3)K activity in liver, we predicted that RBP4 might still alter hepatic regulation of glucose homeostasis, as a dissociation has been found between insulin action on PI(3)K activity and on hepatic glucose production^{24,25}. Several gluconeogenic enzymes, including

phosphoenolpyruvate kinase (Pepck), are regulated by dietary retinol deficiency and/or retinoic acid treatment²⁶. We found that *Pepck* (or *Pck1*) expression was elevated 41% ($P < 0.01$) in the liver of RBP4-injected mice in the fasting state compared with vehicle-injected controls (Fig. 5i). These data indicate that hepatic PEPCK is regulated, either directly or indirectly, by circulating RBP4.

To determine whether RBP4 directly regulates glucose production in hepatocytes, we treated H4IIE rat hepatoma cells with purified recombinant mouse RBP4. H4IIE cells show physiologically appropriate responses to stimuli that regulate hepatic glucose production *in vivo*, including glucocorticoids and insulin^{27,28}. Overnight treatment of H4IIE cells with RBP4 ($100 \mu\text{g ml}^{-1}$) induced a 94% increase in PEPCK protein (Fig. 5j) and increased basal glucose production (Fig. 5k). Furthermore, RBP4 treatment impaired suppression of hepatic glucose production in response to submaximal and maximal insulin concentrations (Fig. 5k). Thus, RBP4 can act directly to induce PEPCK expression, increase basal glucose production, and reduce insulin action to suppress glucose production in hepatocytes.

Discussion

GLUT4 expression is decreased in adipocytes in nearly all insulin-resistant states in humans and rodents², but the mechanism by which this contributes to systemic insulin resistance has not been clear, as adipose tissue contributes relatively little to total body glucose disposal. It now seems that elevated serum RBP4 might be a mechanistic link by which downregulation of GLUT4 in adipocytes contributes to the development or worsening of systemic insulin resistance. We find that RBP4 elevation is a widespread abnormality in insulin-resistant states of various aetiologies. Serum levels and/or urinary excretion of RBP4 have previously been reported to be elevated in humans with type 2 diabetes, but no causal relationship was suggested^{29,30}. Notably, regions near the *RBP4* locus on human chromosome 10q have been linked to increased risk for type 2 diabetes in two different populations^{31,32}.

Rbp4 mRNA is selectively increased in adipose tissue, but not in liver, of insulin-resistant adipose-*Glut4*^{-/-} mice, and rosiglitazone reduces *Rbp4* mRNA in adipose tissue, but not in liver. Although hepatocytes are regarded as the principal source of circulating RBP4 under normal conditions, adipose tissue has the second highest expression level (~ 20 – 40% of levels in liver)³³. As much as 20% of total body retinol may be stored in adipose tissue³³. Within adipose tissue, RBP4 is expressed almost exclusively in adipocytes, in a differentiation-dependent manner³⁴. Furthermore, adipocytes release retinol *in vitro*³⁴. Thus, adipose tissue potentially functions like liver to store and mobilize retinol bound to RBP4. Our data demonstrate that changes in adipocyte-derived RBP4 can have systemic effects on insulin sensitivity and glucose homeostasis.

Although serum RBP4 was consistently elevated in all of the insulin-resistant models we studied, adipose tissue *Rbp4* mRNA (expressed as transcript per microgram of adipose tissue RNA) was increased in some but not all models. For example, increased mRNA levels are seen in *Hsd11b1* transgenic mice (L. Oksanen and J. S. Flier, personal communication), a model of the metabolic syndrome, which is consistent with the results in adipose-*Glut4*^{-/-} mice (Fig. 1). *Rbp4* mRNA levels expressed per adipocyte are also elevated in ob/ob and db/db mice (W. S. Blazer, personal communication). In seeming contrast, we found that *Rbp4* mRNA levels per microgram of adipose tissue RNA were decreased by 40–50% in adipose tissue of both ob/ob mice and mice on a high-fat diet. This discrepancy between *Rbp4* mRNA expressed per adipocyte or per microgram of adipose tissue RNA is probably due to the fact that in rodents with marked hyperinsulinaemia, total RNA content per adipocyte increases up to 4.7-fold³⁵. Elevation of serum RBP4 in insulin-resistant states might be a consequence of increased expression, expanded fat mass, altered secretion and/or altered clearance from the circulation.

RBP4 might act through retinol-dependent or retinol-independent mechanisms. Retinol-dependent mechanisms by which RBP4 may influence insulin action include, but are not limited to, increased production or altered tissue metabolism of retinoic acid isomers, the active forms of retinol that interact with retinoic acid receptors (RARs) and retinoic acid-X receptors (RXRs) to regulate gene transcription³⁶. Consistent with a retinoid-dependent mechanism, we found increased expression of *Pepck*, a retinoid-regulated gene, in the liver of mice injected with RBP4 (Fig. 5i) and in cultured hepatocytes treated with RBP4 (Fig. 5j). We also found modest increases in the expression of several other retinoic acid-responsive genes in muscle of *RBP4*-Tg mice: RAR β 2 increased by 20%, stearoyl-CoA desaturase-1 increased by 37%, and acetyl CoA carboxylase- β showed a 27% increase ($P < 0.05$ for all). Furthermore, the expression of cellular retinoic acid binding protein-1 (*Crabp1*), which plays a role in regulating cellular retinoic acid levels, is increased in muscle of *RBP4*-Tg mice (40%) and mice fed a high-fat diet (3.2-fold). However, the relationship between retinoic acid and insulin resistance is complex, as certain retinoic acid isomers cause insulin resistance and the metabolic syndrome in humans^{37,38}, whereas others activate the RXR-PPAR γ heterodimer and increase insulin sensitivity³⁹. Future studies will determine whether RBP4 elevations in insulin-resistant states result in increased tissue retinol content or altered production of specific retinoic acid isomers.

RBP4 might also cause insulin resistance through a retinol-independent mechanism. Evidence suggests that RBP4 binds with high affinity and high specificity to cell surface receptors^{40,41}. Megalin/gp320, the only RBP4 receptor identified to date in peripheral tissues and a non-specific receptor for macromolecular complexes⁴², binds RBP4 with low affinity ($K_d \sim 2 \mu\text{M}$). A high-affinity receptor for RBP4 has not been identified. RBP4 might transport and deliver other lipophilic molecules in addition to retinol, as shown by the fact that RBP4 binds a wide range of other retinoids *in vitro*⁴³. Finally, RBP4 could also modulate transthyretin function⁴⁴. However, serum transthyretin levels were not altered in *Rbp4* knockout mice⁹, *RBP4*-overexpressing mice¹⁹, several mouse models of insulin resistance (not shown) or in humans with type 2 diabetes²⁹.

Rbp4^{-/-} knockout mice have lower levels of serum free fatty acids (see Supplementary Table 2), which might contribute to their improved insulin sensitivity. However, free fatty acids are not elevated in insulin-resistant *RBP4*-Tg mice (see Supplementary Table 2), in RBP4-injected mice (control $0.446 \pm 0.073 \text{ mM}$; RBP4-injected $0.437 \pm 0.034 \text{ mM}$) or in adipose-*Glut4*^{-/-} mice⁵. Free fatty acid levels are unchanged by fenretinide treatment of mice on a high-fat diet (data not shown). Thus, altering the circulating levels of free fatty acids is probably not the principal mechanism by which RBP4 regulates insulin sensitivity.

A finding with high clinical significance in our study is that normalization of serum RBP4 by fenretinide treatment leads to improved insulin action and glucose tolerance in insulin-resistant obese mice. Increased excretion of RBP4 is probably the main action by which fenretinide reverses insulin resistance in obese rodents. Fenretinide and its metabolites lack a terminal carboxyl group, an essential feature of active retinoids, and fenretinide does not activate RXR isoforms *in vitro*^{45,46}, indicating that the insulin-sensitizing effects of fenretinide are different from those of selective RXR agonists (that is, retinoids)⁴⁷. Although there may be additional mechanisms by which fenretinide improves insulin-glucose homeostasis, our findings establish that the RBP4-transthyretin interaction is a new target for the development of drugs to combat insulin resistance and type 2 diabetes.

METHODS

Mice. Mice with adipose-specific deletion of *Glut4* (adipose-*Glut4*^{-/-}) were generated by Cre/loxP gene targeting⁵. Transgenic mice overexpressing *GLUT4* selectively in adipocytes (adipose-*GLUT4*-Tg)⁴, *RBP4*-Tg¹⁹ and *Rbp4*^{-/-} mice⁹ have been described. Male *RBP4*-overexpressing mice were bred with female

C57BL mice (Taconic) to generate *RBP4*-overexpressing mice (*RBP4*-Tg). Male and female *Rbp4*^{+/-} mice were bred to obtain *Rbp4*^{-/-} mice. As female *RBP4*-Tg and *Rbp4*^{-/-} mice had inconsistent phenotypes, only male mice were used for this study.

Mice were fed a standard chow diet (Formulab 5008, Labdiet 5053) or high-fat diet (55% fat calories) (Harlan-Teklad 93075). These diets contained sufficient amounts of vitamin A ($15\text{--}25 \text{ IU g}^{-1}$). The minimum dietary retinol content required to maintain normal vitamin A status in mice is $2.5 \text{ IU g}^{-1} \text{ diet}^{48}$.

RNA, microarray and Taqman. Between six and nine mice from each of four genotypes were studied using a total of 12 microarray chips: aP2-Cre transgenic mice (controls for adipose-*Glut4*^{-/-} mice), adipose-*Glut4*^{-/-} mice; FVB mice (controls for adipose-*GLUT4*-Tg mice) and adipose-*GLUT4*-Tg mice. Total RNA from epididymal adipose tissue was extracted using the RNeasy Mini Kit from Qiagen. Affymetrix gene chip hybridization and analysis were performed at the Genomics Core Facility of the Beth Israel Deaconess Medical Center. Mouse *Rbp4* mRNA was quantified using real-time PCR (see Supplementary Methods).

Serum RBP4 measurement. Serum or plasma was diluted 20 times in a standard detergent-containing buffer⁵, and proteins were separated by 15% SDS-PAGE and transferred to nitrocellulose membranes. Mouse and human RBP4 proteins were detected using anti-rat⁹ or anti-human (DAKO) RBP4 polyclonal antisera, respectively. The anti-human antibody also recognizes mouse RBP4 but with lower affinity.

Measurement of metabolic parameters, insulin and glucose tolerance tests. Plasma glucose, insulin, free fatty acids, leptin, and adiponectin were measured, and insulin and glucose tolerance tests were performed as described^{5,13} (see Supplementary Methods). Resistin was measured using luminex (Linco).

Recombinant RBP4 preparation and injection. Human and mouse RBP4 were expressed in *E. coli* and purified as described^{49,50}. Recombinant RBP4 protein was completely pure, determined by total protein staining of SDS-PAGE. RBP4 bound retinol stoichiometrically and interacted normally with purified transthyretin. Endotoxin was removed by sequential affinity adsorption to Endotrap matrix (ProfosAG) and Detoxigel (Pierce). Purified RBP4 protein was dialyzed in a buffer containing 10 mM HEPES buffer, 100 mM NaCl, stored frozen at stock concentrations of $7\text{--}8 \text{ mg ml}^{-1}$ and protected from exposure to light. The dialysate solution not containing RBP4 was stored for use as a vehicle control for *in vivo* experiments.

In acute pharmacodynamic studies, stock human RBP4 was diluted in dialysate (5 mg ml^{-1} concentration) and $100 \mu\text{l}$ was injected intraperitoneally (i.p.) into 12-week-old male FVB mice. For chronic administration, stock RBP4 was further diluted in dialysate and 8-week-old FVB mice were injected with $100 \mu\text{l}$ of 0.875 mg ml^{-1} RBP4 solution at 8:00 and 15:00, and with $100 \mu\text{l}$ of 1.25 mg ml^{-1} solution at 22:00. Control mice were injected with an equal volume of dialysate vehicle control solution. Endotoxin levels, measured by Limulus amoebocyte assay (Cambrex/BioWhittaker), were less than 0.01 endotoxin units per microlitre for both the RBP4 and vehicle control solutions, which is less than the ambient endotoxin levels of reverse-osmosis double-deionized water (Millipore).

Fenretinide treatment. Three-week-old male FVB mice received chow diet, high-fat diet or high-fat diet supplemented with 0.04% fenretinide. Fenretinide gel capsules were emptied and added directly to the fat-soluble vitamin component of the high-fat diet in the Harlan-Teklad laboratory at the time of diet preparation. Light exposure was minimized during diet preparation. The diet was stored in the dark at 4°C and replaced in mouse cages at 2–3-day intervals.

Signal transduction and *Pepck* mRNA expression. Mice were fasted for 16–18 h, injected intravenously with saline or insulin (10 U kg^{-1} body weight) and killed 3 min after injection. PI(3)K activity was measured in phosphotyrosine immunoprecipitates as described⁵. Western blotting was performed using phosphospecific antibodies to IRS-1 tyrosine-612 (BioSource) and IR tyrosine-972 (Biomol). *Pepck* mRNA was measured by northern blotting as described¹³.

Hepatocyte glucose production and *Pepck* immunoblotting. Glucose production in H4IIE rat hepatoma cells was measured as described²⁸, except that cells were incubated in medium with reduced serum (0.2% FBS) for 18 h prior to the experiment. Lysates were prepared as described¹³. Western blotting was performed using an anti-rat *Pepck* polyclonal antibody (gift from D. Granner).

Statistical analysis. All values are given as means $\pm$ s.e.m. Differences between two groups were assessed using unpaired two-tailed *t*-tests. Data involving more than two groups were assessed by analysis of variance (ANOVA). Glucose and insulin tolerance tests were assessed by repeated measures ANOVA using Statview Software (BrainPower).

Received 29 March; accepted 3 May 2005.

- Despres, J. P. et al. Hyperinsulinemia as an independent risk factor for ischemic heart disease. *N. Engl. J. Med.* **334**, 952–957 (1996).

2. Shepherd, P. R. & Kahn, B. B. Glucose transporters and insulin action—implications for insulin resistance and diabetes mellitus. *N. Engl. J. Med.* **341**, 248–257 (1999).
3. DeFronzo, R. A. Pathogenesis of type 2 diabetes: metabolic and molecular implications for identifying diabetes genes. *Diabetes Rev.* **5**, 171–269 (1997).
4. Shepherd, P. R. *et al.* Adipose cell hyperplasia and enhanced glucose disposal in transgenic mice overexpressing GLUT4 selectively in adipose tissue. *J. Biol. Chem.* **268**, 22243–22246 (1993).
5. Abel, E. D. *et al.* Adipose-selective targeting of the GLUT4 gene impairs insulin action in muscle and liver. *Nature* **409**, 729–733 (2001).
6. Tozzo, E., Gnudi, L. & Kahn, B. B. Amelioration of insulin resistance in streptozotocin diabetic mice by transgenic overexpression of GLUT4 driven by an adipose-specific promoter. *Endocrinology* **138**, 1604–1611 (1997).
7. Carvalho, E., Kotani, K., Peroni, O. & Kahn, B. B. Adipose-specific overexpression of GLUT4 reverses insulin resistance and diabetes in mice lacking GLUT4 selectively in muscle. *Am. J. Physiol. Endocrinol. Metab.* doi:10.1152/ajpendo.0016.2005 (2005).
8. Kershaw, E. E. & Flier, J. S. Adipose tissue as an endocrine organ. *J. Clin. Endocrinol. Metab.* **89**, 2548–2556 (2004).
9. Quadro, L. *et al.* Impaired retinal function and vitamin A availability in mice lacking retinol-binding protein. *EMBO J.* **18**, 4633–4644 (1999).
10. Minokoshi, Y., Kahn, C. R. & Kahn, B. B. Tissue-specific ablation of the GLUT4 glucose transporter or the insulin receptor challenges assumptions about insulin action and glucose homeostasis. *J. Biol. Chem.* **278**, 33609–33612 (2003).
11. Kitamura, T., Kahn, C. R. & Accili, D. Insulin receptor knockout mice. *Annu. Rev. Physiol.* **65**, 313–332 (2003).
12. Blaner, W. S. Retinol-binding protein: the serum transport protein for vitamin A. *Endocr. Rev.* **10**, 308–316 (1989).
13. Kotani, K., Peroni, O. D., Minokoshi, Y., Boss, O. & Kahn, B. B. GLUT4 glucose transporter deficiency increases hepatic lipid production and peripheral lipid utilization. *J. Clin. Invest.* **114**, 1666–1675 (2004).
14. Masuzaki, H. *et al.* A transgenic model of visceral obesity and the metabolic syndrome. *Science* **294**, 2166–2170 (2001).
15. Minokoshi, Y. *et al.* AMP-kinase regulates food intake by responding to hormonal and nutrient signals in the hypothalamus. *Nature* **428**, 569–574 (2004).
16. Nikouline, S. E. *et al.* Potential role of glycogen synthase kinase-3 in skeletal muscle insulin resistance of type 2 diabetes. *Diabetes* **49**, 263–271 (2000).
17. Lee, C. H., Olson, P. & Evans, R. M. Minireview: lipid metabolism, metabolic diseases, and peroxisome proliferator-activated receptors. *Endocrinology* **144**, 2201–2207 (2003).
18. Kotani, K., Kim, Y. B., Peroni, O., Mundt, A. & Kahn, B. B. Rosiglitazone treatment normalizes glucose tolerance in adipose-specific GLUT4 knockout mice but renders muscle-specific GLUT4 knockout more diabetic [abstract]. *Diabetes* **50**, A274 (2001).
19. Quadro, L. *et al.* Muscle expression of human retinol-binding protein (RBP). Suppression of the visual defect of RBP knockout mice. *J. Biol. Chem.* **277**, 30191–30197 (2002).
20. Naylor, H. M. & Newcomer, M. E. The structure of human retinol-binding protein (RBP) with its carrier protein transthyretin reveals an interaction with the carboxy terminus of RBP. *Biochemistry* **38**, 2647–2653 (1999).
21. Kato, M., Kato, K., Blaner, W. S., Chertow, B. S. & Goodman, D. S. Plasma and cellular retinoid-binding proteins and transthyretin (prealbumin) are all localized in the islets of Langerhans in the rat. *Proc. Natl Acad. Sci. USA* **82**, 2488–2492 (1985).
22. Malpeli, G., Folli, C. & Berni, R. Retinoid binding to retinol-binding protein and the interference with the interaction with transthyretin. *Biochim. Biophys. Acta* **1294**, 48–54 (1996).
23. Mothe, I. & Van Obberghen, E. Phosphorylation of insulin receptor substrate-1 on multiple serine residues, 612, 632, 662, and 731, modulates insulin action. *J. Biol. Chem.* **271**, 11222–11227 (1996).
24. Anai, M. *et al.* Enhanced insulin-stimulated activation of phosphatidylinositol 3-kinase in the liver of high-fat-fed rats. *Diabetes* **48**, 158–169 (1999).
25. Pagliassotti, M. J., Kang, J., Thresher, J. S., Sung, C. K. & Bizeau, M. E. Elevated basal PI 3-kinase activity and reduced insulin signalling in sucrose-induced hepatic insulin resistance. *Am. J. Physiol. Endocrinol. Metab.* **282**, E170–E176 (2002).
26. Shin, D. J., Odom, D. P., Scribner, K. B., Ghoshal, S. & McGrane, M. M. Retinoid regulation of the phosphoenolpyruvate carboxykinase gene in liver. *Mol. Cell. Endocrinol.* **195**, 39–54 (2002).
27. Kahn, C. R., Lauris, V., Koch, S., Crettaz, M. & Granner, D. K. Acute and chronic regulation of phosphoenolpyruvate carboxykinase mRNA by insulin and glucose. *Mol. Endocrinol.* **3**, 840–845 (1989).
28. Wang, J. C., Stafford, J. M., Scott, D. K., Sutherland, C. & Granner, D. K. The molecular physiology of hepatic nuclear factor 3 in the regulation of gluconeogenesis. *J. Biol. Chem.* **275**, 14717–14721 (2000).
29. Basualdo, C. G., Wein, E. E. & Basu, T. K. Vitamin A (retinol) status of First Nation adults with non-insulin-dependent diabetes mellitus. *J. Am. Coll. Nutr.* **16**, 39–45 (1997).
30. Abahusain, M. A., Wright, J., Dickerson, J. W. & de Vol, E. B. Retinol, alpha-tocopherol and carotenoids in diabetes. *Eur. J. Clin. Nutr.* **53**, 630–635 (1999).
31. Meigs, J. B., Panhuysen, C. I., Myers, R. H., Wilson, P. W. & Cupples, L. A. A genome-wide scan for loci linked to plasma levels of glucose and HbA_{1c} in a community-based sample of Caucasian pedigrees: The Framingham Offspring Study. *Diabetes* **51**, 833–840 (2002).
32. Duggirala, R. *et al.* Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *Am. J. Hum. Genet.* **64**, 1127–1140 (1999).
33. Tsutsumi, C. *et al.* Retinoids and retinoid-binding protein expression in rat adipocytes. *J. Biol. Chem.* **267**, 1805–1810 (1992).
34. Zovich, D. C. *et al.* Differentiation-dependent expression of retinoid-binding proteins in BFC-1β adipocytes. *J. Biol. Chem.* **267**, 13884–13889 (1992).
35. Pedersen, O., Kahn, C. R. & Kahn, B. B. Divergent regulation of the Glut 1 and Glut 4 glucose transporters in isolated adipocytes from Zucker rats. *J. Clin. Invest.* **89**, 1964–1973 (1992).
36. Chambon, P. A decade of molecular biology of retinoic acid receptors. *FASEB J.* **10**, 940–954 (1996).
37. Koistinen, H. A. *et al.* Dyslipidemia and a reversible decrease in insulin sensitivity induced by therapy with 13-*cis*-retinoic acid. *Diabetes Metab. Res. Rev.* **17**, 391–395 (2001).
38. Rodondi, N. *et al.* High risk for hyperlipidemia and the metabolic syndrome after an episode of hypertriglyceridemia during 13-*cis* retinoic acid therapy for acne: a pharmacogenetic study. *Ann. Intern. Med.* **136**, 582–589 (2002).
39. Klierer, S. A., Xu, H. E., Lambert, M. H. & Willson, T. M. Peroxisome proliferator-activated receptors: from genes to physiology. *Recent Prog. Horm. Res.* **56**, 239–263 (2001).
40. Sivaprasadarao, A. & Findlay, J. B. The interaction of retinol-binding protein with its plasma-membrane receptor. *Biochem. J.* **255**, 561–569 (1988).
41. Matarese, V. & Lodish, H. F. Specific uptake of retinol-binding protein by variant F9 cell lines. *J. Biol. Chem.* **268**, 18859–18865 (1993).
42. Christensen, E. I. *et al.* Evidence for an essential role of megalin in transepithelial transport of retinol. *J. Am. Soc. Nephrol.* **10**, 685–695 (1999).
43. Berni, R., Clerici, M., Malpeli, G., Clerici, L. & Formelli, F. Retinoids: *in vitro* interaction with retinol-binding protein and influence on plasma retinol. *FASEB J.* **7**, 1179–1184 (1993).
44. Monaco, H. L. The transthyretin-retinol-binding protein complex. *Biochim. Biophys. Acta* **1482**, 65–72 (2000).
45. Sheikh, M. S. *et al.* N-(4-hydroxyphenyl)retinamide (4-HPR)-mediated biological actions involve retinoid receptor-independent pathways in human breast carcinoma. *Carcinogenesis* **16**, 2477–2486 (1995).
46. Um, S. J. *et al.* Antiproliferative mechanism of retinoid derivatives in ovarian cancer cells. *Cancer Lett.* **174**, 127–134 (2001).
47. Shen, Q., Cline, G. W., Shulman, G. I., Leibowitz, M. D. & Davies, P. J. Effects of retinoids on glucose transport and insulin-mediated signalling in skeletal muscles of diabetic (db/db) mice. *J. Biol. Chem.* **279**, 19721–19731 (2004).
48. Subcommittee on Laboratory Animal Nutrition, Committee on Animal Nutrition, Board on Agriculture, National Research Council. *Nutrient Requirements of Laboratory Animals* 4th revised edn, 92 (National Academies Press, Washington DC, 1995).
49. Wang, T. T., Lewis, K. C. & Phang, J. M. Production of human plasma retinol-binding protein in *Escherichia coli*. *Gene* **133**, 291–294 (1993).
50. Xie, Y., Lashuel, H. A., Mirov, G. J., Dikler, S. & Kelly, J. W. Recombinant human retinol-binding protein refolding, native disulfide formation, and characterization. *Protein Expr. Purif.* **14**, 31–37 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to W. S. Blaner and M. Gottesman for RBP4-overexpressing and *Rbp4* knockout mice; T. Ciaraldi and R. R. Henry for human serum samples and metabolic data; Takeda Pharmaceutical Company Limited for recombinant mouse RBP4 and M. Fujisawa for sharing unpublished data; the BIDMC DNA Array Facility for assisting with the Affimetrix array; V. Petkova for assistance with Taqman; E. Rosen, Y.-B. Kim and Y. Minokoshi for helpful suggestions; and C. Wason for technical help. We thank J. E. Smith and B. Mickelson for advice regarding incorporation of fenretinide into rodent diets. Fenretinide was provided by the R. W. Johnson Pharmaceutical Company with the assistance of the National Cancer Institute's Cancer Therapy Evaluation Program. This work was supported by grants from the NIH (B.B.K. and W. Blaner), Takeda Pharmaceutical Company Limited (B.B.K.) and the American Diabetes Association (B.B.K.). T.E.G. and J.M.Z. were supported by NIH career awards, F.P. by a Swiss National Science Foundation fellowship, and N.M. by an American Heart Association fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to B.B.K. (bkahn@bidmc.harvard.edu).

Extreme collisions between planetesimals as the origin of warm dust around a Sun-like star

Inseok Song¹, B. Zuckerman², Alycia J. Weinberger³ & E. E. Becklin²

The slow but persistent collisions between asteroids in our Solar System generate a tenuous cloud of dust known as the zodiacal light (because of the light the dust reflects). In the young Solar System, such collisions were more common and the dust production rate should have been many times larger¹. Yet copious dust in the zodiacal region around stars much younger than the Sun has rarely been found². Dust is known to orbit around several hundred main-sequence stars³, but this dust is cold and comes from a Kuiper-belt analogous region out beyond the orbit of Neptune. Despite many searches, only a few main-sequence stars reveal warm (>120 K) dust analogous to zodiacal dust near the Earth^{3–5}. Signs of planet formation (in the form of collisions between bodies) in the regions of stars corresponding to the orbits of the terrestrial planets in our Solar System have therefore been elusive. Here we report an exceptionally large amount of warm, small, silicate dust particles around the solar-type star BD+20 307 (HIP 8920, SAO 75016). The composition and quantity of dust could be explained by recent frequent or huge collisions between asteroids or other ‘planetesimals’ whose orbits are being perturbed by a nearby planet.

BD+20 307 was first identified^{6,7} as possessing a prominent infrared excess at both 12 and 25 μm but not at 60 and 100 μm , as measured in 1983 by the Infrared Astronomical Satellite (IRAS). Such an excess indicates the presence of a warm dusty disk heated by the central star. Our observations, taken with the Keck and Gemini North telescopes, further reveal prominent silicate emission features in the spectra of BD+20 307's dust in the 8–13 μm region of the infrared spectrum (Fig. 1). Comparison of the spectra to laboratory-measured mass absorption coefficients of amorphous and crystalline silicates⁸ implies a mixture of the two. The strength and shape of the silicate emission features imply that grains cannot be much bigger than $\sim 3 \mu\text{m}$ in diameter⁹. Polycyclic aromatic hydrocarbon (PAH) molecules can create an emission feature near 11.3 μm that is similar to the small hump in our observed spectra. However, PAHs also have very high colour temperatures with rising emission to shorter wavelengths and prominent features at 8–9 μm ; there are no indications of PAHs in these spectra. The interstellar medium is composed entirely of amorphous grains, but when heated to high temperatures, $\sim 1,000 \text{ K}$, amorphous grains can anneal into crystals¹⁰. Disks around stars older than $\sim 1 \text{ Myr}$ commonly contain a substantial mass in crystals, as do Solar System comets^{11,12}. Detailed compositional analysis needs broader spectral coverage requiring space telescope observations.

For main-sequence stars with dusty disks, the measured dust temperature is an indication of how far from a star the dust is located. To estimate dust temperature and the fraction of the stellar luminosity reradiated by the dust, τ , we fitted optical and infrared measurements out to the K-band (2 μm) with a synthetic stellar

atmosphere spectrum¹³ along with a single-temperature (650 K) dusty blackbody (Fig. 2). The assumption of a single dust temperature is appropriate for a narrow ring of dust. The intense emission between 9 and 25 μm must be coming from some combination of dust continuum emission and silicate emission peaks. Integration of the excess emission from the L-band (3.5 μm) to 25 μm indicates that τ is ~ 0.04 . An infrared excess star surrounded by such massive amounts of warm dust without being accompanied by IRAS-detected cooler dust ($T < 100 \text{ K}$) has not previously been seen. However,

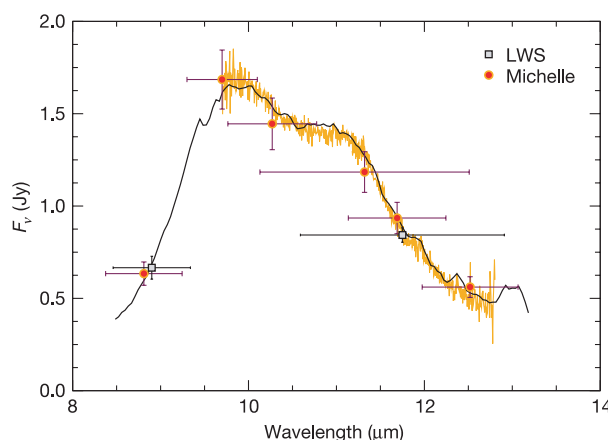


Figure 1 | Infrared spectrum of BD+20 307. The Keck LWS spectrum (black) was obtained on 29 August 2004 with a 0.48-arcsec-wide slit for a total integration time of 440 s. The Gemini Michelle spectra (orange) were obtained with two settings of the medium resolution ($\lambda/\Delta\lambda \approx 900$) grating and the 0.2-arcsec-wide slit. One, centred on 10.5 μm , was taken on 28 September 2004 and another, centred at 12 μm , was taken on 3 October 2004 with 600 s of integration time in each. In all cases, comparison spectra of bright stars observed with the same instrument settings and at nearly the same air mass were used to divide out telluric features. Spectra were normalized to the 11.69- μm flux density measured with Michelle (Table 1). The presence of crystalline silicates is demonstrated by the broad peak at 11.3 μm superimposed on the amorphous, interstellar medium silicate, peak at 10 μm . The grains must be smaller than a few micrometres to emit such a strong feature. PAH features are considerably narrower and the falling flux density toward 8 μm rules out PAHs as the cause of the 11.3- μm peak. The shape of the silicate feature shown here and in Fig. 2 is similar to silicate features seen in many comets^{26,27}, but the BD+20 307 feature is significantly stronger, as is evident in Fig. 2. Some combination of $\sim 3\text{-}\mu\text{m}$ -size amorphous and crystalline olivine and pyroxene grains can reproduce the observed spectra quite well; however, a full compositional analysis needs space telescope data with broader spectral coverage. Flux uncertainties from Table 1 are shown as vertical error bars and 50% power filter passbands are plotted as horizontal bars.

¹Gemini Observatory, 670 North A'ohoku Place, Hilo, Hawaii 96720, USA. ²Department of Physics and Astronomy and Center for Astrobiology, University of California, Los Angeles, Los Angeles, California 90095-1547, USA. ³Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road, NW, Washington DC 20015, USA.

IRAS was not sensitive to modest quantities of cool dust at Sun-like stars 90 pc from the Earth, so a definite understanding of the full range of dust temperatures must await long-wavelength photometry from the Spitzer Space Telescope or ground-based submillimetre observations.

The presence of primordial dust around newly born stars is common, and very young stars frequently show warm dust with high τ (ref. 14). The oldest-known stars with $\tau \geq 0.1$ are TW Hya, Hen 3-600 and HD 98800 in the 8-Myr-old TW Hydrae association¹⁵. To see whether BD+20 307 is such a young star, we analysed an optical echelle spectrum taken with the DuPont 2.5 m telescope at the Las Campanas Observatory. We then used age-dating methods including lithium content in the stellar photosphere, X-ray flux, and space motion; details can be found in ref. 16. The lithium 6708 Å absorption feature (equivalent width 37 ± 8 mÅ) is comparable in strength to that of Ursa Majoris moving group stars with the same effective temperature. Lithium content in a stellar atmosphere is mainly determined by the star's age and effective temperature, so we estimate the age of BD+20 307 to be similar to that of Ursa Majoris moving group stars (~ 300 million years old). The velocities of BD+20 307 toward the centre of our Milky Way galaxy, around the Galactic Center, and perpendicular to the Galactic plane (U, V, W) are calculated from Hipparcos proper motions and parallax, and the radial velocity (-11.0 ± 1.0 km s⁻¹) is measured from the echelle spectrum. The calculated UVW ($-3, -22$ and $+4$ km s⁻¹) differs slightly, but noticeably, from the UVW of known young ≤ 100 -Myr moving groups in the solar neighbourhood (table 7 in ref. 16), but would be consistent with the motion of somewhat older stars. Velocity widths (v) of absorption lines ($v \sin i = 5.5$ km s⁻¹; i is the angle of inclination of the stellar spin axis with respect to the line of sight toward the Earth) of various atomic species also suggest that BD+20 307 is not exceptionally young¹⁷. Furthermore, BD+20 307 was not detected in the ROSAT All-sky X-ray Survey, from which we can deduce an upper limit to the star's X-ray flux. The calculated upper limit of fractional X-ray luminosity ($L_x/L_{\text{bol}} < 10^{-4.9}$) falls

between the loci of Hyades and Pleiades cluster members with similar effective temperatures¹⁶ and is consistent with the age of BD+20 307 determined from lithium and UVW . Using all available data, we estimate an age of ~ 300 Myr for BD+20 307. The star might be older than our estimate, but it is unlikely to be younger than a few hundred million years. Measurement of the soft X-ray flux would better constrain an upper bound to the age.

A τ of 0.04 is unprecedented for a star as old as BD+20 307 and is orders of magnitude larger than the τ values of the dustiest-known main-sequence stars of comparable age. One reason for its relatively large τ is the relative proximity of the dust to BD+20 307; the closer a given grain is to a star, the more effectively it can absorb stellar flux. Surveys of solar-mass stars are incomplete, but a comprehensive study of stars of mass 2.5 times the solar mass shows that by an age of a few hundred million years, excess emission at $24 \mu\text{m}$ is less than twice the level of the stellar photosphere¹⁸. The excess at $24 \mu\text{m}$ of BD+20 307 is 74 times the photosphere. It is of interest whether this dust has remained from the primordial cloud that formed the star, has been regenerated continuously in the collisions of asteroid-like bodies, or was generated recently in a giant collision.

Large, blackbody-like grains at 650 K would be located only ~ 0.25 AU from BD+20 307. However, the small grains responsible for the strong $10\text{-}\mu\text{m}$ silicate feature (Fig. 2) radiate less efficiently than blackbodies and thus would be further (~ 0.4 AU), from the star. Also, cooler dust grains could be present. Thus, in the following discussion, we assume the grains orbit BD+20 307 at a typical distance of 1 AU, but recognize that additional measurements will be needed to confirm or deny this estimate. At ~ 1 AU from the central star, small grains cannot survive long. With $\tau = 0.04$, the collisional grinding timescale is a few hundred years¹⁹ and grains fractured to

Table 1 | Keck and Gemini mid-infrared observations

Wavelength (μm)	Flux (mJy)	Uncertainty (mJy)	Instrument
3.85	291	12	LWS
4.70	246	33	LWS
7.72	208	8	Michelle
8.81	634	63	Michelle
8.90	666	62	LWS
9.70	1,685	160	Michelle
10.27	1,445	140	Michelle
11.32	1,184	110	Michelle
11.69	935	85	Michelle
11.75	843	39	LWS
12.52	561	56	Michelle
17.75	683	65	LWS
18.10	741	70	Michelle
24.50	410	44	LWS
8.5–13.2	–	–	LWS, $\lambda/\Delta\lambda \approx 150$ spectrum
9.7–11.3	–	–	Michelle, $\lambda/\Delta\lambda \approx 900$ spectrum
11.2–12.8	–	–	Michelle, $\lambda/\Delta\lambda \approx 900$ spectrum

In our systematic IRAS search for warm excess among main-sequence stars measured with the Hipparcos astrometric satellite, BD+20 307 stood out prominently. From a fit to the photospheric spectrum (Fig. 2), we derive a temperature of 6,000 K consistent with its G0 spectral type. At a distance of 92 ± 11 pc (~ 300 light years)²⁵, as expected for a ~ 300 -Myr-old Sun-like star, BD+20 307 is 1.8 times more luminous than the Sun. BD+20 307 was observed with two instruments: the Long Wavelength Spectrometer (LWS) at the W. M. Keck Observatory and the mid-infrared imager/spectrometer (Michelle) of the Gemini North Telescope. LWS and Michelle images at six and eight filter bands, respectively, spanning the wavelength range 4–25 μm (the above table and Fig. 2) confirm that the strong IRAS source is indeed coincident with the star. IRAS had a large beam size of ~ 1 arcmin; thus we rule out the possibility that the very unusual infrared excess is due to source confusion within the IRAS field of view. At all wavelengths, the source appears point-like down to the diffraction limit of the Keck and Gemini telescopes (~ 0.4 arcsec). Bright standard stars were observed before and after observations of BD+20 307 and the standard deviation in their photometry was used as an estimate of calibration uncertainty ($\sim 5\%$). Calibration uncertainties and statistical uncertainties are added in quadrature to estimate the above 1σ root-mean-square (r.m.s.) uncertainty of the measurements.

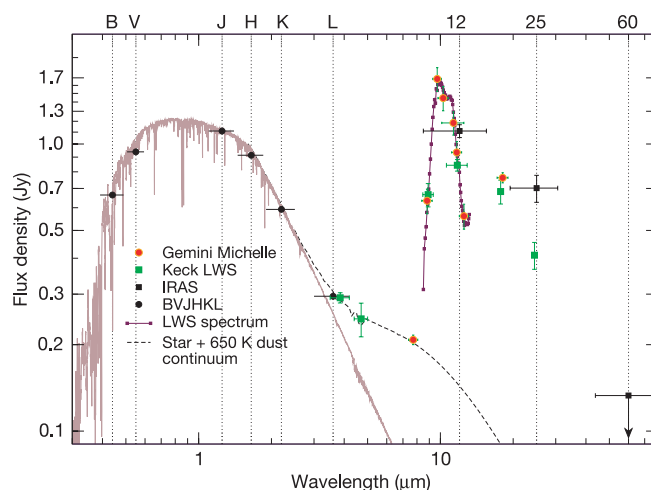


Figure 2 | Spectral energy distribution of BD+20 307. The infrared excess of BD+20 307 produces 3–25- μm flux densities much larger than those expected from the star alone. To estimate the contribution from the star, visual and near-infrared photometric data (0.3–2 μm wavelength) were χ^2 -squared-fitted with a synthetic stellar spectrum (brown line) of 6,000 K and radius 1.25 times that of the Sun. Dust continuum emission was estimated by fitting data shortward of 8 μm with a single blackbody temperature. The sum of the two fits well all data shortward of 8 μm (dashed black curve). The total luminosity of the excess, determined by integrating under the data points between 2 μm (K-band) and 60 μm , is $\sim 4\%$ of the luminosity of the star. Photometric measurements longward of 8 μm are all significantly above the combined model flux from star and dust continuum in part or entirely because of strong silicate emission features (evident, for example, at 10 μm). BV data are from the Hipparcos catalogue²⁵, JHK data are from the 2MASS database²⁸, and the L value is from ref. 6. IRAS fluxes are not colour-corrected because the 12- μm and probably the 25- μm fluxes are dominated by silicate emission and 60 μm and 100 μm are upper limits. Horizontal and vertical error bars have the same meaning as in Fig. 1.

smaller than a few tenths of a micrometre will be instantaneously blown away from the star by the radiation pressure of BD+20 307. The Poynting–Robertson drag timescale for $\sim 1\text{-}\mu\text{m}$ radius particles is only a thousand years²⁰. Therefore, all grains in the system should be second generation, that is, they should be remnants of collisions among larger objects.

Chen and Jura²¹ analysed warm dust near ζ Leporis, a star more massive than BD+20 307 but of comparable age. They postulated a quasi-steady-state model where collisions of parent bodies over the lifetime of ζ Lep require the presence of a region similar to the Sun's asteroid belt, but with 200 times the mass. Because BD+20 307 is much more extreme than ζ Lep ($\tau = 0.04$ and 0.00017 , respectively), a quasi-steady state over the lifetime of BD+20 307 would imply the existence of an asteroid belt $\sim 10,000$ times more massive than the Sun's asteroid belt, that is, a mass of asteroids comparable to Earth's mass.

Consideration of such an extremely large mass of asteroids suggests that what we are witnessing may be the result of recent, perhaps especially violent, collisions; that is, BD+20 307 has not been always so dusty and we now see an unusual and transient event. A similar scenario of such extreme collisions at a much younger (~ 12 Myr) dusty star, β Pictoris, was recently suggested²². IRAS discovered evidence for remnant dust bands in the inner Solar System generated by collisions between Solar System main-belt asteroids²³. Given our above estimates of particle size and semi-major axes, a $\sim 300\text{-km}$ -diameter object, roughly the size of Davida (the fifth-largest asteroid), would recently have been pulverized into tiny particles to create the observed excess in the $3\text{--}25\text{-}\mu\text{m}$ range. When the Sun was as young as BD+20 307, the inner Solar System was undergoing heavy bombardment by asteroids and/or comets, sometimes at a rate 1,000 times larger than at present, but still at levels much below that measured for this star²⁴. Perhaps BD+20 307 is in an even more extreme phase. The contrast between the very short-lived micrometre-sized grains prevalent near BD+20 307 and the much larger $30\text{--}100\text{-}\mu\text{m}$ particles in the Sun's zodiacal cloud further attest to a recent event at BD+20 307. In any case, extreme collisional events occur only rarely at stars—numerous stars of age comparable to that of BD+20 307 and much closer to Earth than 92 pc were observed with the IRAS and Infrared Space Observatory satellites, but nothing similar to BD+20 307 has been observed at any other star. It is of interest to determine whether a particular planetary architecture results in such violent collisions at such late ages.

Received 22 November 2004; accepted 18 May 2005.

1. Kenyon, S. J. & Bromley, B. C. Detecting the dusty debris of terrestrial planet formation. *Astrophys. J.* **602**, L133–L136 (2004).
2. Aumann, H. H. & Probst, R. G. Search for Vega-like nearby stars with 12 micron excess. *Astrophys. J.* **368**, 264–271 (1991).
3. Zuckerman, B. Dusty circumstellar disks. *Annu. Rev. Astron. Astrophys.* **39**, 549–580 (2001).
4. Laureijs, R. J. *et al.* A 25 micron search for Vega-like disks around main-sequence stars with ISO. *Astron. Astrophys.* **387**, 285–293 (2002).
5. Zuckerman, B. & Song, I. Dusty debris disks as signposts of planets: implications for Spitzer Space Telescope. *Astrophys. J.* **603**, 738–743 (2004).
6. Whitelock, P. A. *et al.* South galactic cap G and K stars with infrared excesses. *Mon. Not. R. Astron. Soc.* **250**, 638–643 (1991).

7. Stencel, R. E. & Backman, D. E. A survey for infrared excesses among high galactic latitude SAO stars. *Astrophys. J. Suppl. Ser.* **75**, 905–924 (1991).
8. Koike, C. *et al.* Compositional dependence of infrared absorption spectra of crystalline silicate. *Astron. Astrophys.* **399**, 1101–1107 (2003).
9. Li, A. & Draine, B. T. Infrared emission from interstellar dust. II. The diffuse interstellar medium. *Astrophys. J.* **554**, 778–801 (2001).
10. Hallenbeck, S. L., Nuth, J. A. III & Nelson, R. N. Evolving optical properties of annealing silicate grains: from amorphous condensate to crystalline mineral. *Astrophys. J.* **535**, 247–255 (2000).
11. van Boekel, R. *et al.* The building blocks of planets within the 'terrestrial' region of protoplanetary disks. *Nature* **432**, 479–482 (2004).
12. Hanner, M. S., Lynch, D. K. & Russell, R. W. The 8–13 micron spectra of comets and the composition of silicate grains. *Astrophys. J.* **425**, 274–285 (1994).
13. Hauschildt, P. H., Allard, F. & Baron, E. The NextGen model atmosphere grid for $3000 < T_{\text{eff}} < 10,000$ K. *Astrophys. J.* **512**, 377–385 (1999).
14. Haisch, K. E., Lada, E. A. & Lada, C. J. Disk frequencies and lifetimes in young clusters. *Astrophys. J.* **553**, L153–L156 (2001).
15. Zuckerman, B., Forveille, T. & Kastner, J. H. Inhibition of giant planet formation by rapid gas depletion around young stars. *Nature* **373**, 494–496 (1995).
16. Zuckerman, B. & Song, I. Young stars near the Sun. *Annu. Rev. Astron. Astrophys.* **42**, 685–721 (2004).
17. Cutispoto, G. *et al.* Fast-rotating nearby solar-type stars. *Astron. Astrophys.* **397**, 987–995 (2003).
18. Rieke, G. H. *et al.* Decay of planetary debris disks. *Astrophys. J.* **620**, 1010–1026 (2005).
19. Kenyon, S. J. & Bromley, B. C. Prospects for detection of catastrophic collisions in debris disks. *Astrophys. J.* (in the press) (2005).
20. Wyatt, S. & Whipple, F. The Poynting–Robertson effect on meteor orbits. *Astrophys. J.* **54**, 134–141 (1950).
21. Chen, C. H. & Jura, M. A possible massive asteroid belt around ζ Leporis. *Astrophys. J.* **560**, L171–L174 (2001).
22. Telesco, C. M. *et al.* Mid-infrared images of β Pictoris and the possible role of planetesimal collisions in the central disk. *Nature* **433**, 133–136 (2005).
23. Nesvorný, D., Bottke, W. F., Levison, H. F. & Dones, L. Recent origin of the solar system dust bands. *Astrophys. J.* **591**, 486–497 (2003).
24. Gaidos, E. Observational constraints on late heavy bombardment episodes around young solar analogs. *Astrophys. J.* **510**, L131–L134 (1999).
25. Perryman, M. A. C. *et al.* The Hipparcos Catalogue. *Astron. Astrophys.* **323**, L49–L52 (1997).
26. Haywood, T. L., Hanner, M. S. & Sekanina, Z. Thermal infrared imaging and spectroscopy of comet Hale-Bopp (C/1995 O1). *Astrophys. J.* **538**, 428–455 (2000).
27. Sitko, M. L., Lynch, D. K., Russell, R. W. & Hanner, M. S. 3–1.4 micron spectroscopy of comets C/2002 O4, C/2002 V1, C/2002 X5, C/2002 Y1, and 69P/Taylor. *Astrophys. J.* **612**, 576–587 (2004).
28. Cutri, R. M. *et al.* The 2MASS All-Sky Catalog of Point Sources (University of Massachusetts and IPAC/California Institute of Technology, 2003).

Acknowledgements We thank G. Preston for obtaining the echelle spectrum and S. Fisher for help acquiring Gemini Michelle data. This research was supported by NASA grants to Gemini, CIW and UCLA, and by the UCLA and CIW nodes of the NASA Astrobiology Institute. This paper is based on observations obtained at the Gemini and Keck Observatories. The Gemini Observatory is operated by the Association of Universities for Research in Astronomy, Inc., under a cooperative agreement with the NSF on behalf of the Gemini partnership: the National Science Foundation (United States), the Particle Physics and Astronomy Research Council (United Kingdom), the National Research Council (Canada), CONICYT (Chile), the Australian Research Council (Australia), CNPq (Brazil) and CONICET (Argentina). The Keck Observatory is operated as a scientific partnership among the California Institute of Technology, The University of California and NASA.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to I.S. (song@gemini.edu).

LETTERS

Seismic resurfacing by a single impact on the asteroid 433 Eros

P. C. Thomas¹ & Mark S. Robinson²

Impact cratering creates a wide range of topography on small satellites and asteroids. The population of visible craters evolves with impacts, and because there are no competing endogenic processes to modify the surface, determining the various ways younger craters add to or subtract from the population is a fundamental aspect of small-body geology^{1,2}. Asteroid 433 Eros, the most closely studied small body, has regions of substantially different crater densities^{3–5} that remain unexplained. Here we show that the formation of a relatively young crater (7.6 km in diameter) resulted in the removal of other craters as large as 0.5 km over nearly 40 per cent of the asteroid. Burial by ejecta cannot explain the observed pattern of crater removal. The limitation of reduced crater density to a zone within a particular straight-line distance through the asteroid from the centre of the large crater suggests degradation of the topography by seismic energy⁶ released during the impact. Our observations indicate that the interior of Eros is sufficiently cohesive to transmit seismic energy over many kilometres, and the outer several tens of metres of the asteroid must be composed of relatively non-cohesive material.

Impacts may remove or modify older craters by direct overprinting, erosion, covering with ejecta, or by seismic effects⁶. Ejecta distribution from a single event can be predicted, accounting for the body's shape, density, spin period and models of ejecta velocity distribution^{7–10}. Testing models of seismic modification, either by an initial jolt or by reverberations, has focused on simulating average crater density parameters^{6,11,12}, or possible fracture patterns¹³. The global imaging of asteroid 433 Eros by the NEAR-Shoemaker spacecraft (<4 m per pixel) provides the first opportunity to map small crater (150–500 m) density variations over an entire asteroid, and to test for specific, rather than average, patterns of crater density.

Crater density variations of small (50–300 m) craters have been noted within the large impact features of Psyche, Shoemaker and Himeros^{3–5,14}. A global count of craters >200 m in diameter found a deficiency within and around Shoemaker crater which was attributed to ejecta coverage⁴. We extend this previous work to smaller-diameter craters; the mapped craters include a wide variety of morphologies and degradational stages. Over 99% of the craters are >10 pixels across; so image resolution is not a factor in detecting the craters, and multiple views ensured the inclusion of favourable illumination conditions.

We make a global crater density map by tabulating the number of craters within 2 km of each latitude–longitude grid point of the shape model (2° spacing) and map a standard crater density parameter¹⁵ (R) that is proportional to the number of craters between diameter limits. This sampling radius provides adequate statistics; the maximum number of craters was 50, the minimum 0.

There are two regions of distinctly lower than average crater density: the Shoemaker–Himeros area, and a smaller region extending in two lobes eastward from crater Psyche (Fig. 1). In some places

the transition from high to low crater density can occur over distances comparable to our sampling radius (Fig. 1b, bottom).

Can the regions of low crater density be caused by ejecta from Shoemaker crater as proposed earlier⁴? The transition from heavily cratered to lightly cratered terrain is easily seen in many NEAR images (Fig. 2), but the images do not reveal filling of craters other than interior slumping, nor other clear indications of burial by ejecta. The photogeologic evidence is not definitive, and quantitative tests are more revealing. For the crater sizes affected, the ejecta would have to cover ~390 km² to depths of 50 m; the estimated ejecta volume¹⁶ of 15 km³ could make a uniform covering of ~40 m depth. However, if all the ejecta are within a contiguous area around the crater, depths should decrease rapidly from the rim: lunar ejecta depths vary approximately as Rc^{-3} , where Rc is distance in crater radii¹⁷. Such a distribution would leave much of the area with far less ejecta than required to fill even the smaller craters that we consider.

Although much ejecta may be retained on small objects suffering large impacts^{10,18,19}, rapid rotation and irregular shape typically result in strongly asymmetric ejecta distribution^{10,19,20}. The distribution of blocks >15 m on Eros serves as a proxy for the low-velocity ejecta from Shoemaker¹⁵. The geographic patterns of block density show substantial differences from the patterns of low crater density regions detailed here (Fig. 1d). If the ejecta blocks represent only a small low-velocity fraction, then the volumetric problem becomes even more severe because much of the ejecta would have escaped. If it is a substantial fraction of the ejecta, then the highly asymmetric pattern and its extent show that Shoemaker ejecta are not confined to the areas of low crater density. The ejecta explanation fails because: (1) a realistic concentration of all ejecta fails to cover much of the required area, (2) ejecta trajectory predictions do not explain the pattern east of Psyche, and (3) an actual indicator of ejecta distribution bears no resemblance to the low crater density pattern.

Can the diminution of craters be due to seismic effects of the impact that created Shoemaker crater? Seismic effects of impacts have been studied for multiple events^{12,21,22}, and single, large events^{11,13,23}. The very low gravity on small asteroids is a major factor in suggestions of greater efficiencies of seismically induced regolith motion compared to the Moon or other objects^{12,21,22}.

We have compared crater densities to the straight-line distance through the body of Eros from the centre of Shoemaker crater (14° S, 334° W) in two ways: by geographic pattern (Fig. 1c) and by summary statistics (Fig. 3). On the side of Eros containing Shoemaker, the region of low crater density is confined, with minor exceptions, to the area within a ~9 km straight-line distance of the crater centre. On the side opposite Shoemaker, the distinctive 'wings' of low crater density area are crudely mimicked by the distance function east of Psyche (Fig. 1b, c), but not to the southwest. Statistically, the averaged crater densities (Fig. 3) demonstrate the importance of Shoemaker in the crater density of Eros. Considering the sampling size, areas within a straight-line distance of ~9 km of

¹Center for Radiophysics and Space Research, Cornell University, Ithaca, New York 14853, USA. ²Center for Planetary Sciences, Northwestern University, Evanston, Illinois, USA.

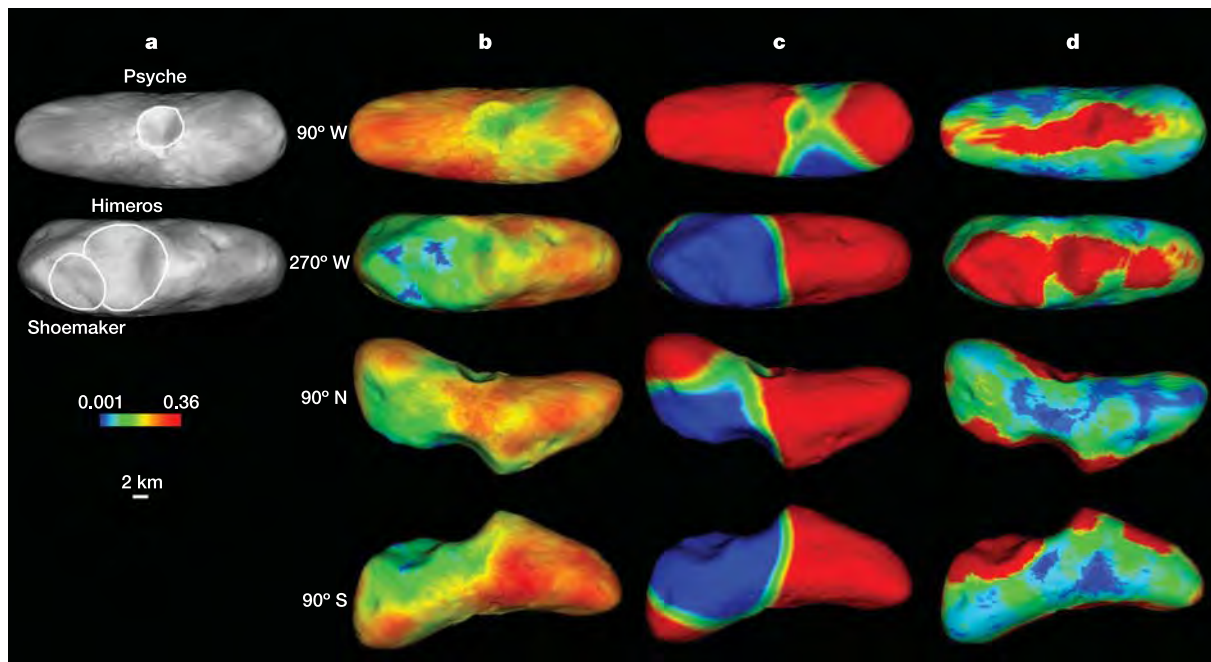


Figure 1 | Crater densities and other surface characteristics of Eros. **a**, Key to locations and scales. The colour scale bar is for **b** only. **b**, Crater densities on Eros. Data binned between diameters of 0.177 to 1 km, within a radius of 2 km of each 2×2 grid point. Grid point spacing varies with latitude and radius, but is always less than 2 km. Values plotted are scaled to the log of the R -value¹⁵ range indicated in the colour scale. $R = [(D_a D_b)^{3/2}] * [N_a / (D_b - D_a)]$, where D_a and D_b are crater diameter ranges, $D_b > D_a$ and N_a is the number of craters per unit area. **c**, Colour-coding of straight-line distance through the asteroid to the centre of crater

Shoemaker. Red, >10.5 km, violet, <8.5 km distance. The area >10.5 km mimics some distinctive parts of the crater density pattern, including the area east of crater Psyche (top row). **d**, Colour-coding of volume of large ejecta blocks per unit area. Block data are from a previous study¹⁶, with scaling to distinguish variations outside of Shoemaker crater, which contains most of Eros's block volume. These blocks are proxies for low-velocity ejecta from the crater, and indicate that the low crater density pattern is not the result of burial by ejecta from Shoemaker. Shoemaker crater is termed Charlois Regio by the International Astronomical Union.

the centre of Shoemaker have reduced crater density, involving craters as large as 0.5 km in diameter (Fig. 3b). This simple relation and the matching of peculiarities of the geography of the low crater density and the distance function (Fig. 1d) strongly suggest that the crater degradation was caused by seismic energy release from Shoemaker crater's formation.

The magnitude of distant seismic effects may depend upon the energy released upon impact, distance from the source, attenuation properties of the intervening materials^{12,21,24,25} and the nature of the material where the modification occurs^{12,21,22}. Additionally, such factors as total energy deposited, accelerations, slopes, and local gravity may be important^{11,21,22}. Our data suggest that whatever the mechanisms, on Eros the total effect varies as some simple function of straight-line distance through the asteroid (d) from the source, with a substantial erasure of craters up to 0.5 km occurring up to 9 km from the source crater. If the primary factor in surface modification is deposition of energy per unit area, the effect should scale as d^{-2} , times any attenuation factor²¹, which for a particular value of d would present a surface pattern indistinguishable from d^{-1} . Modelling of cumulative seismic effects¹² includes decrease of energy density (figure S8 in ref. 12) away from the source that is at least as rapid as d^{-2} . The primary geographical variable to consider in addition to distance may be local gravity (g), because the lower the gravity, the farther loose material might be thrown by a particular surface acceleration^{12,21,22}. A parameter of $d^2 g$ causes little change from the plots of R as a function of d (Fig. 3c) because gravity changes only modestly over Eros²⁰ (0.24 – 0.56 cm s⁻²), and the d^2 term dominates the result. Slopes make less of an organized effect because their direction and magnitude vary rapidly on local scales.

Are there other craters that have seismically altered the surface? Psyche, at 5.4 km across, and Himeros, at 10.2 km across, are obvious candidates. We assume the seismic energy, and the morphologic

effects, vary as D^3 (D is crater diameter; the exponent could be slightly larger¹⁷). Using a d^{-2} and D^3 scaling, the amount of degradation observed at 9 km from the centre of Shoemaker would occur at 5.4 km from Psyche: 2 crater radii. The crater density as a

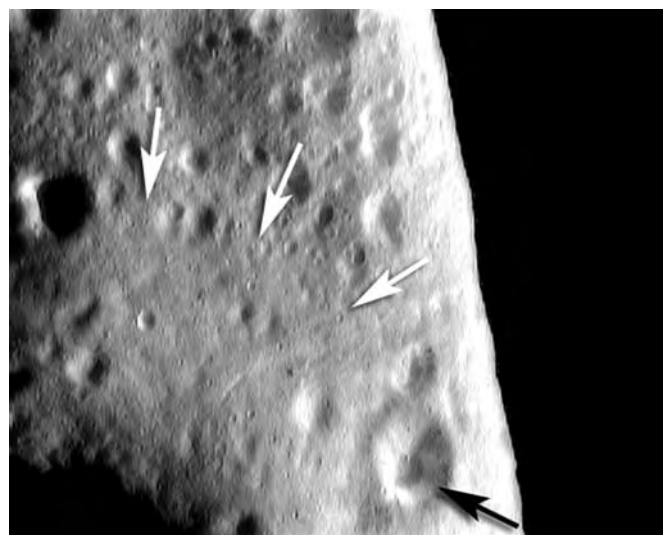


Figure 2 | Transition between regions of different crater density. Bright arrows show some points along the transition of surface smoothness; lower crater density towards bottom, higher crater density towards top. The centre of illuminated surface at 42° S, 242° W is south of the rim of Himeros crater, which is to the lower left of the image. Dulcinea crater (dark arrow) is 1.4 km in diameter, and is centred at 75° S, 272° W. NEAR image M01494091674.

function of distance from Psyche (but excluding the segment possibly affected by Shoemaker) shows only a suggestion of an effect within about 2 crater radii (Fig. 3d). This result may reflect Psyche's greater age than Shoemaker¹⁶, which would allow the accumulation of a crater population near Psyche after its formation. For Himeros, equivalent crater degradation should extend to 14 km from the centre of the crater. Because of Himeros' central location on Eros, its formation should have removed the great majority of craters ≤ 0.5 km in diameter. High crater density areas about the north side of Himeros, indicating that Himeros is very old, and that its low crater density is in large measure the result of Shoemaker's formation.

What is the physical alteration implied by the change in crater density? Fresh craters 0.177 to ~ 0.5 km in diameter have depths of 35–100 m (refs 5, 26). Thus, we are concerned with modifications of a few tens of metres over nearly half the asteroid. Destruction of craters requires the rearrangement of material locally, not just acceleration of a whole landscape. If acting upon coherent rock, the seismic motions may have little effect on the local topography¹⁷. The loss of such craters from shaking implies that the upper tens of metres are in large part non-cohesive material^{11,21}. Images of Eros show sharp features on parts of some degraded craters, and polygonal crater shapes indicative of some structural control²⁷, so crater modification on Eros is not entirely smoothing of sandpile-like material. The loss of ~ 0.5 -km-diameter craters 9 km from a 7.6-km-diameter crater is very close to the model predictions¹¹ for jolt modification of Ida (diameter of 30 km) and Gaspra (diameter of 12 km).

Our results show a clear relation of local crater modification to a single impact event on a small irregular body, an effect not previously documented. We cannot yet disentangle effects arising from an initial jolt or in subsequent reverberations. Globally averaged seismic energy density from many events has been used¹² to predict crater density averaged over the entire surface. Such modelling indicates that although effects at a particular location may include reverberation, for a particular impact the amount of energy deposited decreases sharply with distance from the source (figure S8 in ref. 12), which is consistent with our findings, and consistent with a relatively small role for reverberations close to large events. The morphologic effects observed on Eros can provide benchmarks for further hydrocode and other computations that investigate the release and transmission of impact energy in small bodies.

Because a particular level of effect occurs a constant distance from the source, seismic energy transmission seems to be close to isotropic, despite possible planar structures^{27,28}. The mismatch of the crater-density/distance relation south of Psyche crater may be an area subject to anisotropy effects. The surface modification is a strong indicator that several tens of metres of relatively loose material (regolith) cover most of Eros. Because craters on Eros do not show morphologies indicative of excavation through significant near-surface physical discontinuities⁵, the regolith is probably not strongly layered and its change with depth is gradual and complex²⁹; this conclusion is in agreement with other modelling¹². The results show that detailed knowledge of small crater populations is a valuable tool for obtaining information about the subsurface properties of small bodies.

Received 28 February; accepted 24 May 2005.

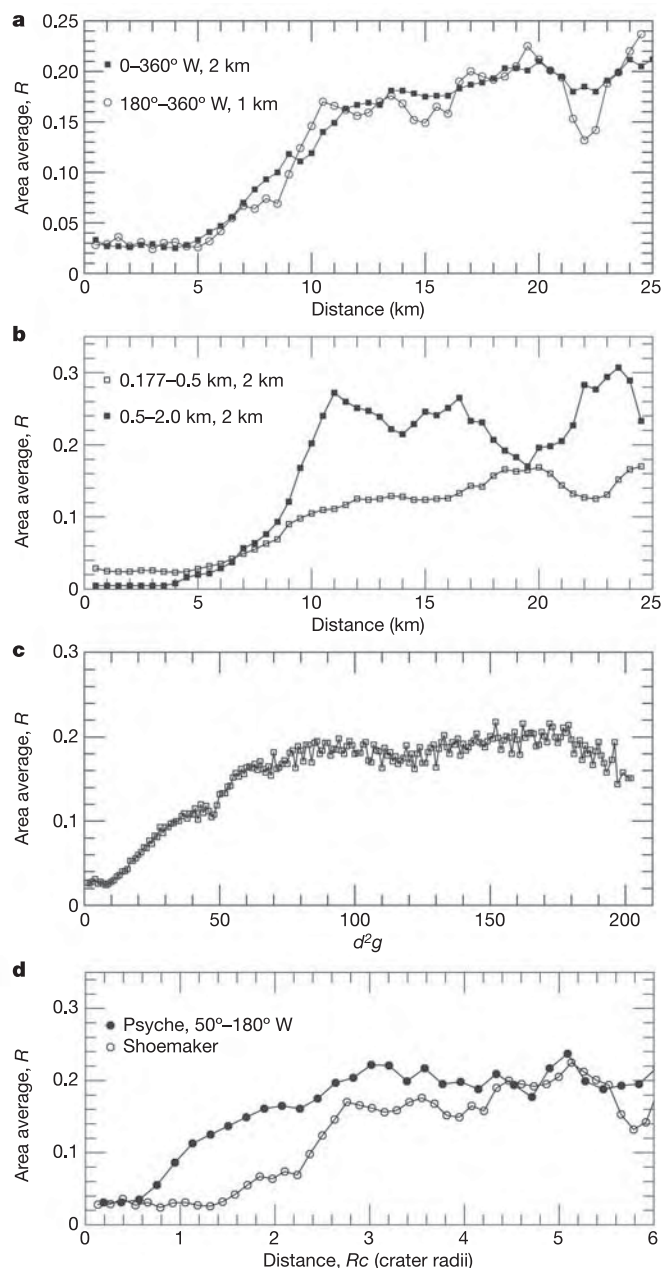


Figure 3 | Summary statistics of crater densities. Crater density as a function of straight-line distance through the body of the asteroid from the centre of Shoemaker crater. **a**, Craters 0.177–1 km in diameter. Filled squares are for entire global area, sampling radius of 2 km. Open circles are for 2 km sampling radius of the 180°–360° W longitude region. **b**, Data for 180°–360° W area, 2 km distance sampling; size ranges as indicated. The large bin size strongly influences the R -values for the 0.5–2 km range sample. Both size ranges of craters show depletion within 10 km of the centre of Shoemaker. **c**, Area-averaged R -values as a function d^2g of straight-line distance d , and local gravity g , a possible scaling for seismic energy effects. **d**, Area-averaged R as a function of distance in crater radii from Shoemaker and from Psyche. The Psyche data use 1-km-radius sampling areas because of the small radius (2.7 km) of Psyche. Additionally, the Psyche data are restricted to the west of 50° W, to avoid the complications of the 'wings' of low crater density best associated with Shoemaker (Fig. 1b).

1. Chapman, C. R. in *Asteroids III* (eds Bottke, W. F., Cellino, A., Paolicchi, P. & Binzel, R.) 315–330 (Univ. Arizona Press, Tucson, 2002).
2. Sullivan, R., Thomas, P. C., Murchie, S. L. & Robinson, M. S. in *Asteroids III* (eds Bottke, W. F., Cellino, A., Paolicchi, P. & Binzel, R.) 331–350 (Univ. Arizona Press, Tucson, 2002).
3. Bethoud, M., Thomas, P. & Veverka, J. Eros: Crater densities in three major impact features. *Bull. Am. Astron. Soc.* **33**, 1149 (2001).
4. Bethoud, M., Veverka, J. & Thomas, P. C. Crater distribution and erasure on asteroid 433 Eros. *Icarus* (submitted).
5. Robinson, M. S., Thomas, P. C., Veverka, J., Murchie, S. L. & Wilcox, B. B. The geology of Eros. *Meteorit. Planet. Sci.* **37**, 1651–1684 (2002).
6. Greenberg, R., Nolan, M. C., Bottke, W. F. & Kolvoord, R. Collisional history of Gaspra. *Icarus* **107**, 84–97 (1994).
7. Dobrovolskis, A. R. & Burns, J. A. Life near the Roche limit: Behavior of ejecta from satellites close to planets. *Icarus* **42**, 422–441 (1980).
8. Thomas, P. C. Ejecta emplacement on the martian satellites. *Icarus* **131**, 78–106 (1998).
9. Korycansky, D. G. & Asphaug, E. Simulations of impact ejecta and regolith accumulation on asteroid Eros. *Icarus* **171**, 110–119 (2004).

10. Geissler, P. *et al.* Erosion and ejecta reaccrusion on 243 Ida and its moon. *Icarus* **120**, 140–157 (1996).
11. Greenberg, R. *et al.* Collisional and dynamical history of Ida. *Icarus* **120**, 106–118 (1996).
12. Richardson, J. E., Melosh, H. J. & Greenberg, R. Impact-induced seismic activity on asteroid 433 Eros: A surface modification process. *Science* **306**, 1526–1529 (2004).
13. Asphaug, E. *et al.* Mechanical and geological effects of impact cratering on Ida. *Icarus* **120**, 158–184 (1996).
14. Chapman, C. R. *et al.* Impact history of Eros: Craters and boulder. *Icarus* **155**, 104–118 (2002).
15. Arvidson, R. E. *et al.* Crater Analysis Techniques Working Group Standard techniques for presentation and analysis of crater size frequency data. *Icarus* **37**, 467–474 (1979).
16. Thomas, P. C. *et al.* Shoemaker crater as the source of most ejecta blocks on the asteroid 433 Eros. *Nature* **413**, 394–396 (2001).
17. Melosh, H. J. *Impact Cratering: A Geologic Process* (Oxford Univ. Press, New York, 1989).
18. Asphaug, E. & Melosh, H. J. The Stickney impact of Phobos: A dynamical model. *Icarus* **101**, 144–164 (1993).
19. Nolan, M. C., Asphaug, E., Melosh, H. J. & Greenberg, R. Impact craters on asteroids: Does gravity or strength control their size? *Icarus* **124**, 359–371 (1996).
20. Thomas, P. C. *et al.* Eros: shape, topography and slope processes. *Icarus* **155**, 18–37 (2002).
21. Houston, W. N., Moriwaki, Y. & Chang, C. S. Downslope movement of lunar soil and rock caused by meteoroid impact. *Proc. Lunar Sci. Conf.* **4**, 2425–2435 (1973).
22. Cheng, A. F., Izenberg, N., Chapman, C. R. & Zuber, M. T. Ponded deposits on asteroid 433 Eros. *Meteorit. Planet. Sci.* **37**, 1095–1105 (2002).
23. Nolan, M. C., Asphaug, E., Greenberg, R. & Melosh, H. J. Impacts on asteroids: fragmentation, regolith transport, and disruption. *Icarus* **152**, 1–15 (2001).
24. Housen, K. R., Holsapple, K. A. & Voss, M. E. Compaction as the origin of the unusual craters on the asteroid Mathilde. *Nature* **402**, 155–157 (1999).
25. Davis, D. R. The collisional history of asteroid 253 Mathilde. *Icarus* **140**, 49–52 (1999).
26. Barnouin-Jha, O. *et al.* Preliminary impact crater dimensions on 433 Eros from the NEAR laser rangefinder and imager. *Lunar Planet. Sci.* **32**, 1786 [CD-ROM] (Lunar and Planetary Institute, Houston, Texas, 2001).
27. Prockter, L. M. *et al.* Surface expressions of structural features on Eros. *Icarus* **155**, 75–93 (2002).
28. Thomas, P. C., Prockter, L., Robinson, M., Joseph, J. & Veverka, J. Global structure of asteroid 433 Eros. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL014599 (2002).
29. Veverka, J. *et al.* Imaging of small-scale features on 433 Eros from NEAR: Evidence for a complex regolith. *Science* **292**, 484–488 (2001).

Acknowledgements This work was supported in part by the NASA Discovery Data Analysis Program. We are grateful to B. Carcich, to K. Consroe for technical help, and to M. Berthoud and J. Veverka for discussions. We also thank E. Asphaug and R. Greenberg.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.T. (thomas@baritone.astro.cornell.edu).

LETTERS

Massively parallel manipulation of single cells and microparticles using optical images

Pei Yu Chiou¹, Aaron T. Ohta¹ & Ming C. Wu¹

The ability to manipulate biological cells and micrometre-scale particles plays an important role in many biological and colloidal science applications. However, conventional manipulation techniques—including optical tweezers^{1–6}, electrokinetic forces (electrophoresis^{7,8}, dielectrophoresis⁹, travelling-wave dielectrophoresis^{10,11}), magnetic tweezers^{12,13}, acoustic traps¹⁴ and hydrodynamic flows^{15–17}—cannot achieve high resolution and high throughput at the same time. Optical tweezers offer high resolution for trapping single particles, but have a limited manipulation area owing to tight focusing requirements; on the other hand, electrokinetic forces and other mechanisms provide high throughput, but lack the flexibility or the spatial resolution necessary for controlling individual cells. Here we present an optical image-driven dielectrophoresis technique that permits high-resolution patterning of electric fields on a photoconductive surface for manipulating single particles. It requires 100,000 times less optical intensity than optical tweezers. Using an incoherent light source (a light-emitting diode or a halogen lamp) and a digital micromirror spatial light modulator, we have demonstrated parallel manipulation of 15,000 particle traps on a $1.3 \times 1.0 \text{ mm}^2$ area. With direct optical imaging control, multiple manipulation functions are combined to achieve complex, multi-step manipulation protocols.

Light-patterned electrodes have been widely used in xerography, which was invented in 1942¹⁸. This concept was recently applied to patterning of colloidal structures^{19,20}; optically-induced electrophoresis was used to attract charged particles onto indium tin oxide (ITO)¹⁹ and semiconductor²⁰ surfaces. However, none of the previous literature has shown the capability of single-particle manipulation. Our optoelectronic tweezers (OET) utilize direct optical images to create high-resolution dielectrophoresis (DEP) electrodes for the parallel manipulation of single particles. DEP force results from the interaction of the induced dipoles in particles subjected to a non-uniform electric field⁹. The magnitude of the force depends on the electric field gradient and the polarizability of the particle, which is dependent on the dielectric properties of the particle and the surrounding medium. DEP is a well established technique, and has been widely used to manipulate micrometre and submicrometre particles as well as biological cells^{21,22}. Travelling-wave DEP is particularly attractive for high-throughput cell manipulation without external liquid pumping^{10,11}. The travelling electric field produced by a multi-phase alternating current (a.c.) bias on a parallel array of electrodes levitates and transports many particles simultaneously. However, travelling-wave DEP cannot resolve individual particles. Recently, a programmable DEP manipulator with an individually addressable two-dimensional electrode array has been realized using complementary metal-oxide-semiconductor (CMOS) integrated circuit technology²³: parallel manipulation of a large number ($\sim 10,000$) of cells was demonstrated. The CMOS DEP manipulator has two potential drawbacks: first, the need of on-chip integrated circuits increases the cost of the chip, making it less

attractive for disposable applications; and second, the trap density ($\sim 400 \text{ sites mm}^{-2}$) is also limited by the size of the control circuits.

Figure 1 illustrates the OET device structure used in our experiments. The liquid containing the cells or particles of interest is sandwiched between a upper transparent, conductive ITO-coated glass, and a lower photoconductive surface, which consists of multiple featureless layers of ITO-coated glass, an n+ hydrogenated amorphous silicon (a-Si:H) layer, an undoped a-Si:H layer, and a silicon nitride layer. These two surfaces are biased with an a.c. signal,

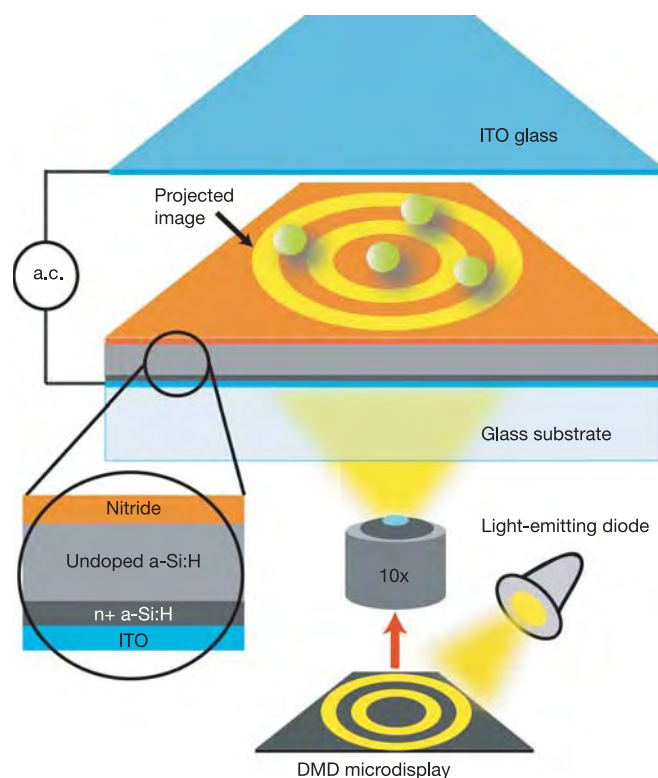


Figure 1 | Device structure used in optoelectronic tweezers. Liquid that contains microscopic particles is sandwiched between the top indium tin oxide (ITO) glass and the bottom photosensitive surface consisting of ITO-coated glass topped with multiple featureless layers: 50 nm of heavily doped hydrogenated amorphous silicon (a-Si:H), 1 μm of undoped a-Si:H, and 20 nm of silicon nitride. The top and bottom surfaces are biased with an a.c. electric signal. The illumination source is an light-emitting diode operating at a wavelength of 625 nm (Lumileds, Luxeon Star/O). The optical images shown on the digital micromirror display (DMD) are focused onto the photosensitive surface and create the non-uniform electric field for DEP manipulation.

¹Department of Electrical Engineering and Computer Sciences, and Berkeley Sensor and Actuator Centre (BSAC), University of California at Berkeley, California 94720, USA.

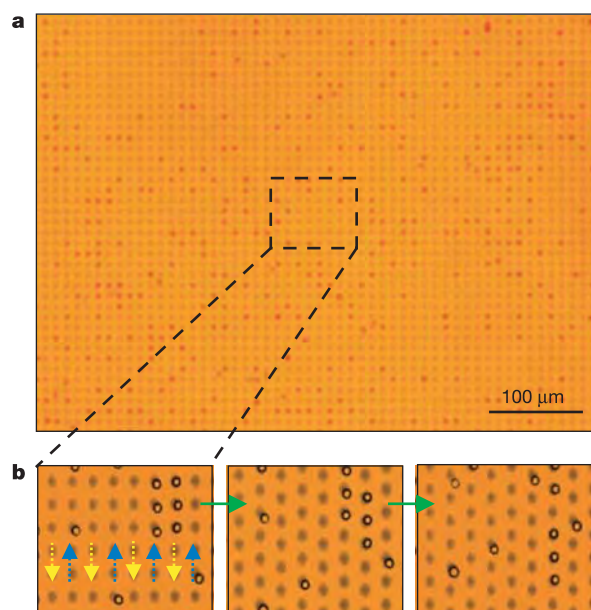


Figure 2 | Massively parallel manipulation of single particles. **a**, 15,000 particle traps are created across a $1.3 \text{ mm} \times 1.0 \text{ mm}$ area. The $4.5\text{-}\mu\text{m}$ -diameter polystyrene beads experiencing negative DEP forces are trapped in the darker circular areas. Each trap has a diameter of $4.5 \text{ }\mu\text{m}$, which is adjusted to fit a single particle. **b**, Parallel transportation of single particles. Three snapshots from the video show the particle motion in part of the manipulation area. The trapped particles in two adjacent columns move in opposite directions, as indicated by the blue and yellow arrows.

10 V peak-to-peak. When projected light illuminates the photoconductive layer, it turns on the virtual electrodes, creating non-uniform electric fields and enabling particle manipulation via DEP forces. These featureless layers are made without photolithography in fabrication, making the device inexpensive and attractive for disposable applications. The OET-based optical manipulation has two operational modes, positive OET and negative OET, as a result of the DEP forces induced for actuation. Particles can be attracted by or repelled from the illuminated area, depending on the a.c. electric field frequency and the particle's internal and surface dielectric properties.

Thanks to the photoconductive gain, the minimum optical intensity required to turn on a virtual electrode is $10 \text{ nW } \mu\text{m}^{-2}$, which is 100,000 times lower than that used in optical tweezers. This opens up the possibility of using incoherent optical images to control the DEP forces over a large area. The optical images are created by combining a light-emitting diode and a digital micromirror spatial light modulator (Texas Instruments, $1,024 \times 768$ pixels, $13.68 \text{ }\mu\text{m} \times 13.68 \text{ }\mu\text{m}$ pixel size). The pattern is imaged onto the photoconductive surface through a $10\times$ objective. The resulting pixel size of the virtual electrode is $1.52 \text{ }\mu\text{m}$. The illumination source is a red light-emitting diode (625 nm wavelength) with a 1-mW output power (measured after the objective lens), which is sufficient to actuate 40,000 pixels. Tight focusing is not required for OET, and the optical manipulation area can be magnified by choosing an appropriate objective lens. Using a $10\times$ objective, the manipulation area ($1.3 \text{ mm} \times 1.0 \text{ mm}$) is 500 times larger than that of optical tweezers.

Patterning high-resolution virtual electrodes is critical for achieving single-particle manipulation. OET has higher resolution than the optically-induced electrophoretic methods reported previously^{19,20}. The minimum size of the virtual electrode is limited by the lateral diffusion length of the photogenerated carriers in the photoconductor, as well as the optical diffraction of the objective lens. The large number of electronic defect states in undoped a-Si:H results in a short ambipolar electron diffusion length of less than 115 nm (ref. 24). The ultimate virtual electrode resolution is thus determined by the optical diffraction limit. In addition, the induced OET force is

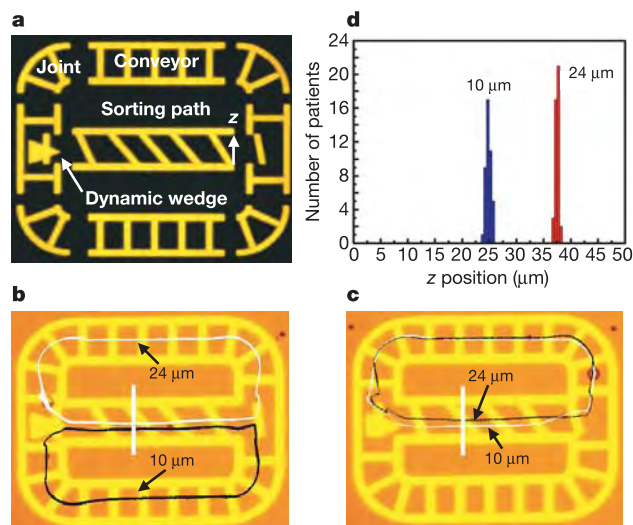


Figure 3 | An example of an integrated virtual optical machine.

a, Integration of virtual components, including an optical sorter path, conveyors, joints and a wedge. The motion of different components is synchronized. **b, c**, Two polystyrene particles with sizes of $10 \text{ }\mu\text{m}$ and $24 \text{ }\mu\text{m}$ pass through the sorter path and are fractionated in the z direction owing to the asymmetrical optical patterns. The particle trajectories can be switched at the end of the sorter path by the optical wedge. **d**, Optical sorting repeatability test. The white and black loops in **b** and **c** represent the particle traces after 43 cycles. The trace broadening at the white bar has a standard deviation of $0.5 \text{ }\mu\text{m}$ for the $10\text{-}\mu\text{m}$ bead and $0.15 \text{ }\mu\text{m}$ for the $24\text{-}\mu\text{m}$ bead.

proportional to the gradient of the square of the electric field, making it well confined to the local area of the virtual electrodes, which is also a key property for single-particle manipulation. A demonstration of the high-resolution capabilities of OET is the creation of 15,000 DEP traps across an area of $1.3 \times 1.0 \text{ mm}^2$ (Fig. 2). The particles are trapped in the darker circular areas by the induced negative DEP forces, which push the beads into the non-illuminated regions, where the electric field is weaker. The size of each trap is optimized to capture a single $4.5\text{-}\mu\text{m}$ -diameter polystyrene bead. By programming the projected images, these trapped particles can be individually moved in parallel (Fig. 2b, and Supplementary Movie 1). Compared with the programmable CMOS DEP chip²³, the particle trap density

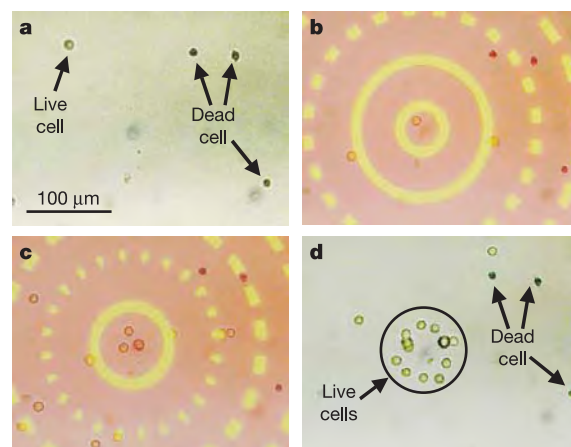


Figure 4 | Selective collection of live cells from a mixture of live and dead cells. **a**, Randomly positioned cells before OET. **b, c**, Cell sorting. The live cells experience positive OET, trapping them in the bright areas, and pulling the live cells into the pattern's centre. The dead cells (stained with Trypan blue dye) leak out through the dark gaps and are not collected. The optical pattern has a yellowish colour, while weak background scattered light results in a pinkish hue in the non-patterned areas. **d**, Sorted cells.

of OET ($11,500 \text{ sites mm}^{-2}$) is 30 times higher, thanks to the high-resolution addressing ability.

Although OET has the capability to pattern high-resolution virtual electrodes, the trapping of submicrometre particles requires a strong electric field gradient to overcome brownian motion. In our current OET, the electric field gradient patterned by the light-emitting diode is not strong enough to trap particles less than $1 \mu\text{m}$ through the DEP mechanism. However, the trapping of particles $1 \mu\text{m}$ or less in diameter does occur when we operate our present device at a low-frequency a.c. bias ($\sim 1 \text{ kHz}$). (See Supplementary Movie 2.) This trapping mechanism is not due to DEP forces, but is caused by optically-patterned electrokinetic flow. Details of this mechanism are still under investigation.

Using direct imaging, sophisticated virtual electrodes can be easily patterned and reconfigured to create dynamic electric field distributions for continuous particle manipulation without the assistance of fluidic flow. Figure 3 shows an example of an integrated optical manipulator that combines the functions of optical conveyors, sorters, wedges and joints. Particles are transported through different functional areas and recycled in this light-patterned circuit, travelling through different paths depending on the position of the wedge divider. Particles with different sizes are fractionated in the lateral z direction as they pass through the sorter path, owing to the asymmetric shape of the light-patterned electric fields. At the end of the sorter path, an optical wedge divides and guides the particles into the two conveyors. The looped optical conveyors recycle the particles back to the sorter input to repeat the process (see Supplementary Movie 3). Figure 3b, c shows that the paths of the fractionated particles can be switched by reconfiguring the tip position of the optical wedge. The trajectories of the particle movement are highly repeatable and accurately defined. Figure 3d shows distribution of the particle position in the middle of the sorter (marked by a white bar) after the particles have passed through the sorter 43 times. The standard deviations of trace broadening are $0.5 \mu\text{m}$ for the $10\text{-}\mu\text{m}$ bead, and $0.15 \mu\text{m}$ for the $24\text{-}\mu\text{m}$ bead. As the magnitude of the DEP force is proportional to the particle volume, the larger particle shows a better confinement in the optically-patterned DEP cages during transportation.

By exploiting the dielectric differences between different particles or cells, DEP techniques have been able to discriminate and sort biological cells that have differences in membrane properties (permeability, capacitance and conductivity), internal conductivity, and size^{21,25,26}. The OET technique not only inherits these DEP advantages, but also provides the capability of addressing each individual cell. We demonstrate the selective concentration of live human B cells from a mixture of live and dead cells in Fig. 4. The cells are suspended in an isotonic buffer medium of 8.5% sucrose and 0.3% dextrose, mixed with a solution of 0.4% Trypan blue dye to check the cell viability, resulting in a conductivity of 10 mS m^{-1} . The applied a.c. signal is 14 V peak-to-peak at a frequency of 120 kHz. The cell membranes of live cells are selectively permeable, and can maintain an ion concentration differential between the intracellular and extracellular environments. Dead cells cannot maintain this differential. When dead cells are suspended in a medium with a low ion concentration, the ions inside the cell membrane are diluted through ion diffusion. This results in a difference between the dielectric properties of live and dead cells²¹. Live cells experience positive OET, and are collected in the centre of the shrinking optical ring pattern by attraction to the illuminated region, while dead cells experience negative OET and are not collected (see Supplementary Movie 4).

Single-cell analysis is an important technique in comprehending many biological mechanisms, as it looks at the spectrum of response of each individual cell under stimulation. In addition to biological applications, the high-resolution electric field patterned on the OET surface could also serve as a dynamic template to guide the crystallization of colloidal structures.

Received 27 March; accepted 11 May 2005.

- Grier, D. G. A revolution in optical manipulation. *Nature* **424**, 810–816 (2003).
- Ashkin, A., Dziedzic, J. M. & Yamane, T. Optical trapping and manipulation of single cells using infrared-laser beams. *Nature* **330**, 769–771 (1987).
- MacDonald, M. P., Spalding, G. C. & Dholakia, K. Microfluidic sorting in an optical lattice. *Nature* **426**, 421–424 (2003).
- Curtis, J. E., Koss, B. A. & Grier, D. G. Dynamic holographic optical tweezers. *Opt. Commun.* **207**, 169–175 (2002).
- McGloin, D., Spalding, G. C., Melville, H., Sibbett, W. & Dholakia, K. Three-dimensional arrays of optical bottle beams. *Opt. Commun.* **225**, 215–222 (2003).
- Garces-Chavez, V., Dholakia, K. & Spalding, G. C. Extended-area optically induced organization of microparticles on a surface. *Appl. Phys. Lett.* **86**, 031106 (2005).
- Kremser, L., Blaas, D. & Kenndler, E. Capillary electrophoresis of biological particles: Viruses, bacteria, and eukaryotic cells. *Electrophoresis* **25**, 2282–2291 (2004).
- Cabrera, C. R. & Yager, P. Continuous concentration of bacteria in a microfluidic flow cell using electrokinetic techniques. *Electrophoresis* **22**, 355–362 (2001).
- Hughes, M. P. Strategies for dielectrophoretic separation in laboratory-on-a-chip systems. *Electrophoresis* **23**, 2569–2582 (2002).
- Pethig, R., Talary, M. S. & Lee, R. S. Enhancing traveling-wave dielectrophoresis with signal superposition. *IEEE Eng. Med. Biol. Mag.* **22**, 43–50 (2003).
- Morgan, H., Green, N. G., Hughes, M. P., Monaghan, W. & Tan, T. C. Large-area travelling-wave dielectrophoresis particle separator. *J. Micromech. Microeng.* **7**, 65–70 (1997).
- Yan, J., Skoko, D. & Marko, J. F. Near-field-magnetic-tweezer manipulation of single DNA molecules. *Phys. Rev. E* **70**, 011905 (2004).
- Lee, H., Purdon, A. M. & Westervelt, R. M. Manipulation of biological cells using a microelectromagnet matrix. *Appl. Phys. Lett.* **85**, 1063–1065 (2004).
- Hertz, H. M. Standing-wave acoustic trap for noninvasive positioning of microparticles. *J. Appl. Phys.* **78**, 4845–4849 (1995).
- Kessler, J. O. Hydrodynamic focusing of motile algal cells. *Nature* **313**, 218–220 (1985).
- Sundararajan, N., Pio, M. S., Lee, L. P. & Berlin, A. A. Three-dimensional hydrodynamic focusing in polydimethylsiloxane (PDMS) microchannels. *J. Microelectromech. Syst.* **13**, 559–567 (2004).
- Lee, G. B., Hwei, B. H. & Huang, G. R. Micromachined pre-focused M x N flow switches for continuous multi-sample injection. *J. Micromech. Microeng.* **11**, 654–661 (2001).
- Pai, D. M. & Springett, B. E. Physics of electrophotography. *Rev. Mod. Phys.* **65**, 163–211 (1993).
- Hayward, R. C., Saville, D. A. & Aksay, I. A. Electrophoretic assembly of colloidal crystals with optically tunable micropatterns. *Nature* **404**, 56–59 (2000).
- Ozkan, M., Bhatia, S. & Esener, S. C. Optical addressing of polymer beads in microdevices. *Sens. Mater.* **14**, 189–197 (2002).
- Gascoyne, P. et al. Microsample preparation by dielectrophoresis: isolation of malaria. *Lab Chip* **2**, 70–75 (2002).
- Krupke, R., Hennrich, F., von Lohneysen, H. & Kappes, M. M. Separation of metallic from semiconducting single-walled carbon nanotubes. *Science* **301**, 344–347 (2003).
- Manaresi, N. et al. A CMOS chip for individual cell manipulation and detection. *IEEE J. Solid-State Circuits* **38**, 2297–2305 (2003).
- Schwarz, R., Wang, F. & Reissner, M. Fermi-level dependence of the ambipolar diffusion length in amorphous-silicon thin-film transistors. *Appl. Phys. Lett.* **63**, 1083–1085 (1993).
- Becker, F. F. et al. Separation of human breast-cancer cells from blood by differential dielectric affinity. *Proc. Natl Acad. Sci. USA* **92**, 860–864 (1995).
- Yang, J., Huang, Y., Wang, X. B., Becker, F. F. & Gascoyne, P. R. C. Differential analysis of human leukocytes by dielectrophoretic field-flow-fractionation. *Biophys. J.* **78**, 2680–2689 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E.R.B. McCabe, U. Bhardwaj, R. Sun and F. Yu at UCLA for providing cultured human B cells for our experiments. We also thank A. Wheeler for technical advice regarding our cell experiments. This project is supported by the Center for Cell Mimetic Space Exploration (CMISE), a NASA University Research, Engineering and Technology Institute (URETI), and the Defense Advanced Research Project Agency (DARPA). P.Y.C. acknowledges support from the Graduate Research and Education in Adaptive Bio-Technology (GREAT) training program. A.T.O. acknowledges support from a National Science Foundation fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.C.W. (wu@eecs.berkeley.edu).

Direct observation of electron dynamics in the attosecond domain

A. Föhlisch¹, P. Feulner², F. Hennies¹, A. Fink², D. Menzel², D. Sanchez-Portal³, P. M. Echenique³ & W. Wurth¹

Dynamical processes are commonly investigated using laser pump–probe experiments, with a pump pulse exciting the system of interest and a second probe pulse tracking its temporal evolution as a function of the delay between the pulses^{1–6}. Because the time resolution attainable in such experiments depends on the temporal definition of the laser pulses, pulse compression to 200 attoseconds (1 as = 10^{-18} s) is a promising recent development. These ultrafast pulses have been fully characterized⁷, and used to directly measure light waves⁸ and electronic relaxation in free atoms^{2–4}. But attosecond pulses can only be realized in the extreme ultraviolet and X-ray regime; in contrast, the optical laser pulses typically used for experiments on complex systems last several femtoseconds (1 fs = 10^{-15} s)^{1,5,6}. Here we monitor the dynamics of ultrafast electron transfer—a process important in photo- and electrochemistry and used in solid-state solar cells, molecular electronics and single-electron devices—on attosecond timescales using core-hole spectroscopy. We push the method, which uses the lifetime of a core electron hole as an internal reference clock for following dynamic processes^{9–19}, into the attosecond regime by focusing on short-lived holes with initial and final states in the same electronic shell. This allows us to show that electron transfer from an adsorbed sulphur atom to a ruthenium surface proceeds in about 320 as.

When studying electron transfer processes in complex systems, it is of equal importance to address the temporal evolution of the electron wave packet and the question of which atomic centre an electron is localized at before charge transfer to the substrate occurs. This atom specific information cannot be provided in pump–probe experiments in the spectral regime of optical transitions. By adapting an element specific synchrotron based soft X-ray spectroscopy method, namely core-hole clock spectroscopy, we can effectively determine on an attosecond timescale electron transfer dynamics originating from an atomically localized state by making use of extremely fast Coster–Kronig decay processes of core-excited states.

The principle of core-hole clock spectroscopy is to take the core-hole lifetime τ as an internal reference clock for the temporal evolution of a dynamic process under investigation^{9–19}. To study charge transfer on the timescale of τ , the dynamics of an electron resonantly excited into an unoccupied state from an atomically localized adsorbate core level (Fig. 1a) is monitored through the autoionization process that accompanies the core-hole decay (Fig. 1b and c). If the initially excited core electron remains in an atomically localized resonance, a linear relation between the energies of the incoming photon and of the outgoing electron in the autoionization is observed (Fig. 1b). This is the so-called Raman autoionization channel at constant binding energy (I). In contrast, if the initial

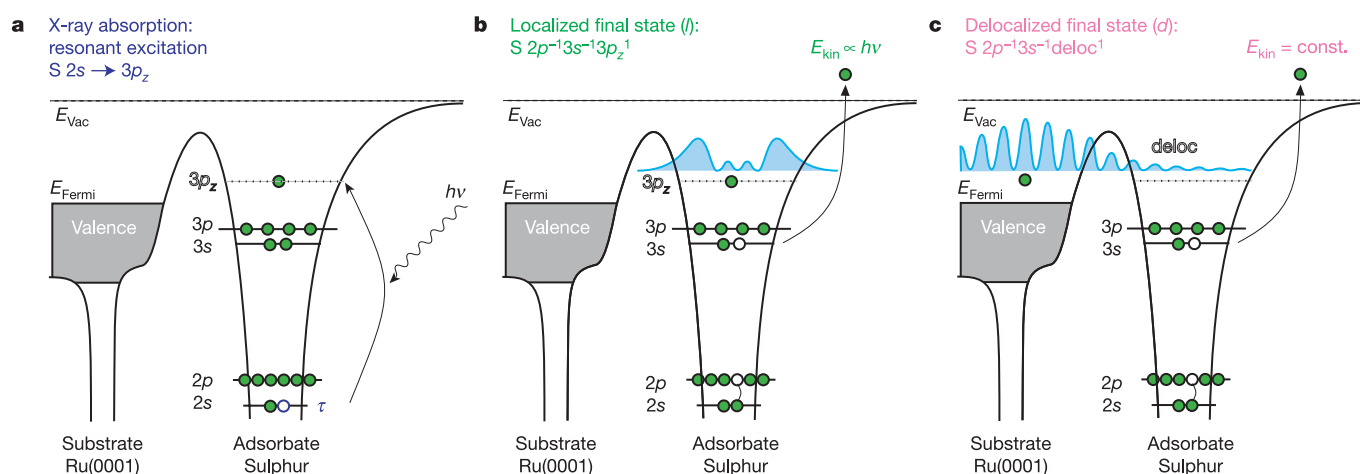


Figure 1 | Core-hole clock spectroscopy—schematic overview. **a**, Initially, a core electron is promoted by resonant excitation from the S 2s level into a bound resonance localized at an adsorbed sulphur atom ($S 2s^{-1} 3p_z^1$) on ruthenium $c(4 \times 2)S/Ru(0001)$ with a core-hole lifetime $\tau = 0.5$ fs. In the autoionization decay processes, Coster–Kronig decay of the S 2s core hole takes place in the presence of this electron, the so-called ‘spectator’ electron,

leading to two different final states. **b**, Localized final state $S 2p^{-1} 3s^{-1} 3p_z^1$: state *l*. The initially excited electron is still localized at the sulphur atom. **c**, Delocalized final state $S 2p^{-1} 3s^{-1} \text{deloc}^1$: state *d*. The initially excited electron has already left the localized resonance. E_{Vac} , vacuum energy; E_{Fermi} , Fermi energy; E_{kin} , kinetic energy.

¹Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, D-22761 Hamburg, Germany. ²Physik Department E20, Technische Universität München, D-85747 Garching, Germany. ³Centro Mixto CSIC-UPV/EHU “Unidad de Física de Materiales”, Donostia International Physics Center (DIPC), and Departamento de Física de Materiales, Universidad del País Vasco, Apdo. 1072, 20080 Donostia-San Sebastián, Spain.

excitation involves an electronic state delocalized over many atomic centres (that is, the excited atomic resonance is coupled to a continuum), we obtain independently of the incident photon energy autoionization at constant kinetic electron energy (Fig. 1c). This is the charge transfer channel of autoionization (*d*). Owing to this different dispersive behaviour, the Raman (*l*) and charge transfer (*d*) channels of autoionization can be spectroscopically separated (Fig. 2a), and the ratio of Raman to charge transfer intensity is related to the degree of atomic localization in the excited state on the timescale τ of the core-hole decay. This can be translated into a dynamic picture of an electron residence time, or alternatively as the charge transfer time, τ_{CT} of electron hopping to the substrate. As spectral intensities are compared, the $\sqrt{N}$ uncertainty (where N is the number of events) allows a statistically significant analysis only as long as the intensities of the spectroscopic channels are less than one order of magnitude apart. Thus a temporal range of charge transfer times between $0.1\tau \leq \tau_{CT} \leq 10\tau$ becomes accessible. The typical core-hole lifetimes of inner shell vacancies lie at oxygen KLL ($\tau = 4$ fs; ref. 20), nitrogen KLL ($\tau = 5$ fs; ref. 20), carbon KLL ($\tau = 6$ fs; ref. 20), and argon $L_3M_{4/5}M_{4/5}$ ($\tau = 6$ fs; refs 13, 18). We note that the core-hole lifetimes depend only weakly on the chemical environment. In a comparison between atomic and molecular systems, variations of the order of roughly 20% have been observed²⁰.

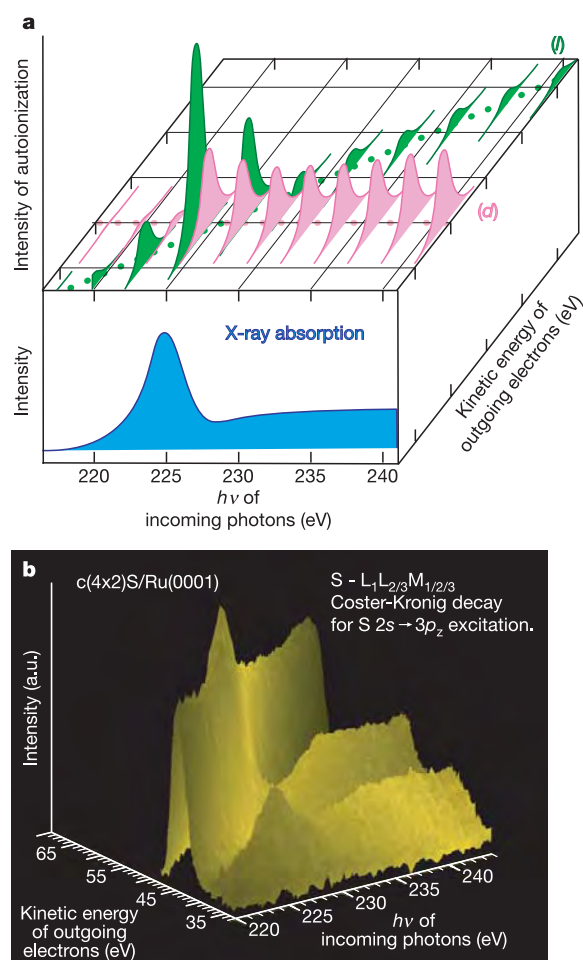


Figure 2 | Core-hole clock spectroscopy—the spectroscopic signatures. **a**, Diagram of the spectroscopic autoionization signatures leading to a localized final state (*l*) with linear dispersion and a delocalized final state (*d*) at constant kinetic energy, and their relation to resonant excitation by X-ray absorption. **b**, Experimental sulphur $L_{1/2/3}M_{1/2/3}$ Coster–Kronig autoionization spectra of $c(4 \times 2)\text{S}/\text{Ru}(0001)$ as a function of incident photon energy.

To access dynamic processes in the attosecond range reliably, shorter core-hole lifetimes are required. Our approach is to perform attosecond charge transfer core-hole clock spectroscopy in the soft X-ray region by monitoring Coster–Kronig autoionization channels

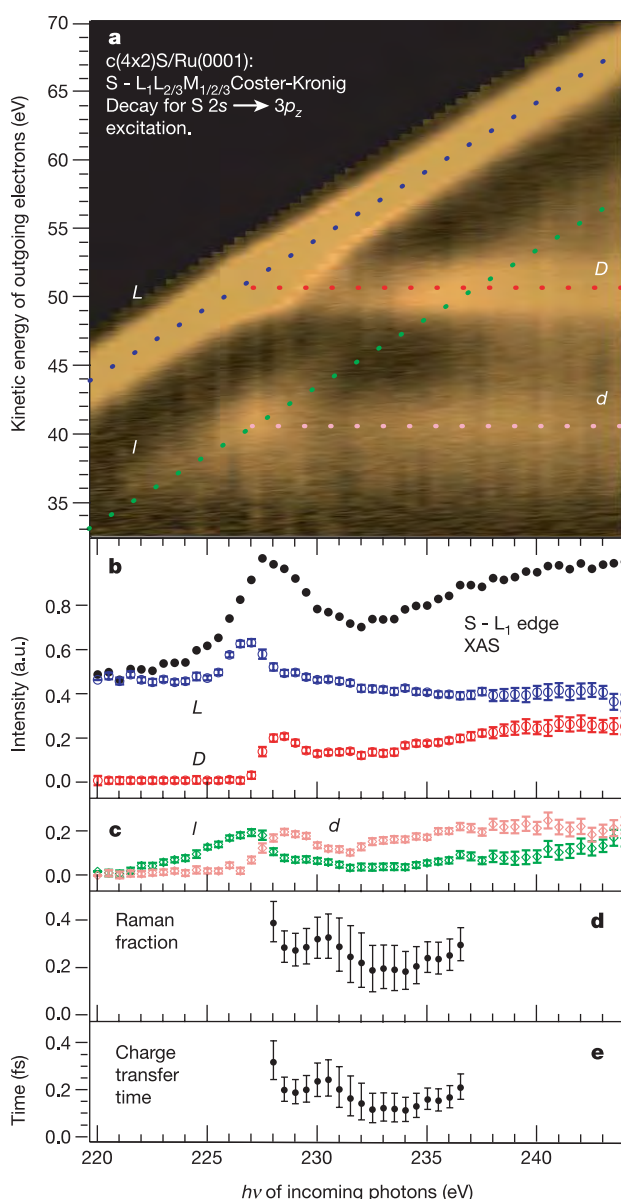
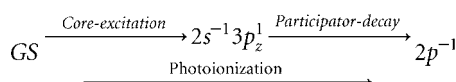


Figure 3 | Quantitative charge transfer analysis of sulphur $L_{1/2/3}M_{1/2/3}$ Coster–Kronig autoionization spectra of $c(4 \times 2)\text{S}/\text{Ru}(0001)$ as a function of photon energy. **a**, Experimental intensities as a function of incoming photon energy and kinetic energy of the outgoing electrons. Lighter colours correspond to higher autoionization intensity. Shown are Raman channels with linear dispersion for localized final states *L* ($2p^{-1}3p^{-1}3p_z^1$) at 170.7 eV binding energy and *l* ($2p^{-1}3s^{-1}3p_z^1$) at 181.7 eV binding energy, and charge transfer channels with delocalized final states *D* ($2p^{-1}3p^{-1}\text{deloc}^1$) at 50.8 eV kinetic energy and *d* ($2p^{-1}3s^{-1}\text{deloc}^1$) at 40.6 eV kinetic energy. **b**, Sum of spectral intensities representing the S- L_1 edge X-ray absorption spectrum. Also shown are separate intensities of the spectral channels (*L*, *D*) from curve fitting with Lorentzians of 3.3 eV FWHM. Error bars show the standard deviation of each fit. **c**, Separate intensities of the spectral channels (*l*, *d*) from curve fitting with Lorentzians of 3.3 eV FWHM. Error bars show the standard deviation of each fit. **d**, Raman fraction $f = l/(l + d)$ as a function of photon energy. Error bars are derived from the standard deviation of the fits (see **c**). **e**, Charge transfer time obtained from the Raman fraction as $\tau_{CT} = \tau f/(1 - f)$ and the S $2s$ core-hole lifetime $\tau = 0.5$ fs. Error bars are derived from the standard deviation of the fits (see **c**).

with attosecond core-hole lifetimes. Here the initial and final state vacancies are in the same electronic shell (same principal quantum number n); the probability for these transitions is higher and the corresponding core-hole lifetimes shorter than in the case of decay processes involving different values of n .

With the $c(4 \times 2)$ -S/Ru(0001) surface^{21,22}, we obtained the S $L_{2/3}M_{1/2/3}$ Coster–Kronig autoionization spectra shown in Fig. 2b as a function of the incoming photon energy $h\nu$ tuned across the $S2s^{-1}3p_z^1$ core level resonance at $h\nu = 227.5$ eV. On resonance, the excited electron's energy lies 1.68 ± 0.1 eV above the Fermi level. We observe resonant enhancement and branching of decay channels at this absorption resonance. The data are converted to a colour-coded plot in Fig. 3a, where higher intensity corresponds to lighter colour. We can directly discern spectral features (l , L) at constant binding energy, which branch into charge transfer spectral features (d , D) with constant kinetic energy. Let us assign the spectral features: starting from the electronic ground state (GS), the $S2s^{-1}3p_z^1$ core-excited state can autoionize through Coster–Kronig channels with and without participation of the excited electron ('participator' and 'spectator' channels, respectively). The participator channel, involving the core-excited $3p_z^1$ in the decay, leads to the spin–orbit split $2p^{-1}$ final state, identical to the main lines of photoionization, with 161.5 eV ($2p_{3/2}^{-1}$) and 162.6 eV ($2p_{1/2}^{-1}$) binding energy. Its spectral features lie outside the range of Figs 2b and 3a.



The spectral features shown in Figs 2b and 3a are thus associated with spectator channels, in particular the $2p^{-1}3s^{-1}3p_z^1$ (l) and

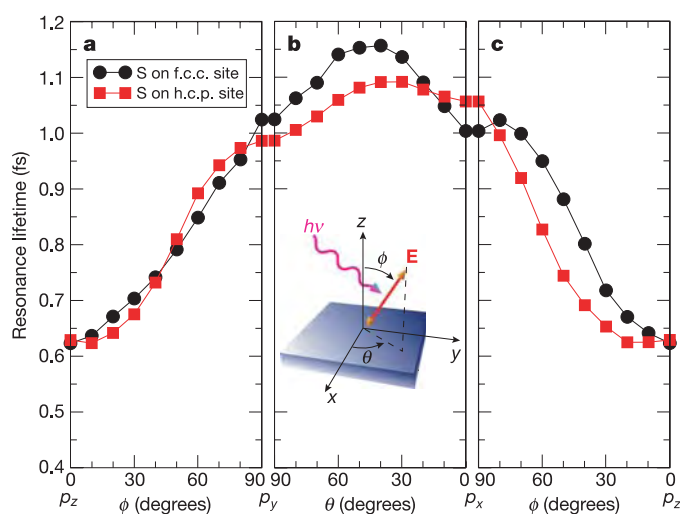
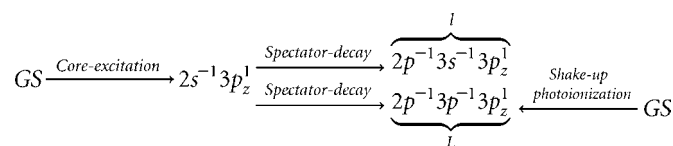


Figure 4 | Theoretical charge transfer time for S in f.c.c. and h.c.p. hollow sites computed as S $3p$ resonance lifetime. Theory predicts a strong dependence of the charge transfer time on the symmetry of the initial wave packet, which translates to a strong dependence on the light polarization. The coordinates x , y , z shown in the diagram in **b** correspond to the crystallographic directions [100], [010], [001]. The circles and squares correspond respectively to sulphur atoms in f.c.c. and h.c.p. sites of the surface. **a**, Resonance lifetime as a function of the angle ϕ of the electric field vector with respect to the surface normal in the y – z plane. Polarization of the synchrotron light along the z axis ($\phi = 0$, the experimental geometry) produces an initial excited state with p_z symmetry. **b**, Resonance lifetime as a function of the angle θ of the electric field vector with respect to the x axis in the x – y plane. In this geometry, p_x and p_y symmetries and combinations of them would be obtained with in-plane polarization. **c**, Resonance lifetime as a function of the angle ϕ of the electric field vector with respect to the surface normal in the x – z plane.

$2p^{-1}3p^{-1}3p_z^1$ (L) final states at 181.7 eV and 170.7 eV binding energy, respectively.



The $2p^{-1}3s^{-1}3p_z^1$ final state l can only be reached via autoionization, whereas the $2p^{-1}3p^{-1}3p_z^1$ final state L can in addition be reached as a photoionization satellite, obeying the monopole selection rule of photoionization shake-ups. Thus, L is present at all photon energies, whereas l is a pure autoionization feature observed only when a core hole has been present. We therefore base all further analysis on the autoionization channel l ($2p^{-1}3s^{-1}3p_z^1$) and the related charge transfer feature d ($2p^{-1}3s^{-1}deloc^1$), which branches off at 40.6 eV constant kinetic energy. The final states l and d are shown schematically in Fig. 1b and c.

To quantify the relative strength of the Raman (l) and charge transfer (d) channels, we performed a curve fit of the spectra at all photon energies with fixed line shapes, where each channel was described phenomenologically by a lorentzian of 3.3 eV full-width at half-maximum (FWHM), and only the intensities were varied. The charge transfer peaks (D , d) were kept at constant kinetic energy and the Raman peaks (L , l) at constant binding energy with varied photon energy. In Fig. 3b and c, these four contributions (D , L , d , l) and the standard deviation of the fit at each photon energy are shown together with their sum. The latter yields the S- L_1 edge X-ray absorption spectrum (XAS) (Fig. 3b) with 3 eV FWHM dominated by the S $2s$ core-hole lifetime of $\tau = 0.5$ fs (ref. 23). From this fit, as a function of photon energy, the relative strength of the Raman contribution expressed as the Raman fraction $f = l/(l + d)$ and the charge transfer time $\tau_{CT} = \tau f/(1 - f)$ using the S $2s$ core-hole life time ($\tau = 0.5$ fs) is derived and displayed in Fig. 3d and e, respectively.

For photon energies below the $S2s^{-1}3p_z^1$ absorption resonance ($h\nu = 227.5$ eV), we observe intensity in the Raman channel l only, as charge transfer below threshold is energetically forbidden, equivalent to an infinitely long charge transfer time. Just above the $S2s^{-1}3p_z^1$ resonance at $h\nu = 228$ eV, we determine $\tau_{CT} = 0.32 \pm 0.09$ fs. For higher photon energies, shorter charge transfer times down to 0.11 ± 0.06 fs at $h\nu = 234$ eV are found. An energy dependence of the electron transfer time has been observed before^{15,18}. This energy dependence is most probably due to the detailed nature of the projected band structure of the substrate, and thus the number and character of the final states available, as well as their overlap with the initial adsorbate state.

We have compared our experimental results with first-principles computations of the charge-transfer dynamics in the $c(4 \times 2)$ -S/Ru(0001) system. The initial electron wave packet is constructed as a linear combination of the S $3p$ orbitals projected onto the unoccupied bands of the combined system, that is, the Ru substrate with the adsorbed sulphur atoms, at the resonance energy. We find that the excitation into a resonance with predominantly $3p_z$ character yields a charge transfer time of 0.63 ± 0.15 fs (Fig. 4a). This corresponds to the state that is excited in the experiment. In comparing experiment to theory, the theoretical time constant confirms that the charge transfer process takes place well below a femtosecond timescale. In particular, the agreement with the experimental value of 0.32 ± 0.09 fs is very satisfactory, taking into account that the core vacancy is not described explicitly in the theoretical ground state calculation; and the theoretical resonance position at ~ 2 eV above the Fermi level is shifted relative to the experimental absorption resonance, which is at 1.68 ± 0.1 eV above the Fermi level. Furthermore, theory predicts a detailed dependence of the charge transfer time on the symmetry of the initial excited state,

which is summarized in Fig. 4. For excitation into $3p_x$ or $3p_y$ -like resonances (in plane) (Fig. 4b) with a smaller overlap to the substrate, a significantly larger charge transfer time of up to 1.15 ± 0.15 fs is calculated. This theoretical result indicates that different polarizations of the light favour different initial excited states, with different symmetries and overlaps with the states of the substrate, thus leading to different transfer times.

The demonstration that soft X-ray spectroscopy can be used as a tool to study the motion of electrons on attosecond timescales opens a possible way to interesting new research areas. Potentially the method is suited to the investigation of electron transfer in complex molecular systems. In such investigations, the ability to excite individual atomic centres (even atoms of the same element in chemically different environments) by exploiting core level shifts should be of particular importance. A second future application could be the investigation of spin-dependent electron transfer processes, which are important in spintronics. Here core level excitation using circularly polarized light could be used to create spin-polarized excitations.

METHODS

Experiments. The experiments were performed at beamline I311, MAX-lab in Lund, Sweden. At 5×10^{-11} torr base pressure, a clean Ru(0001) surface was prepared by cycles of Ar^{2+} -ion sputtering, oxygen-exposure and annealing. The $c(4 \times 2)\text{S}/\text{Ru}(0001)$ surface, with sulphur atoms chemisorbed in hexagonal close packed (h.c.p.) and face-centred cubic (f.c.c.) hollow sites^{21,22}, was prepared by dissociative adsorption of 400 Langmuir H_2S (1 Langmuir = 10^{-6} torr s) at 550 K and annealing to 850 K. The surface quality was checked by core electron spectroscopy (XPS) and low-energy electron diffraction (LEED). At 7° grazing incidence, the electric field vector of the incident radiation was 7° off the surface normal, exciting preferentially into the $\text{S}3p_z$ orbital oriented normal to the surface. The electron spectrometer (Scienta SES 200) was in the polarization plane at 45° to the incident radiation. Narrow bandwidth excitation and high spectral resolution are prerequisites for separating charge-transfer from non-charge-transfer states. Thus the bandwidth of the incident radiation and the ΔE of the electron analyser were both set to 100 meV.

Electronic structure calculations. The density functional calculation of the electronic structure of $c(4 \times 2)\text{S}/\text{Ru}(001)$ has been performed using the SIESTA code^{24,25}. We used a symmetric slab containing 7 Ru layers and the surface geometry known from LEED^{21,22}. Approximately 5 eV below the Fermi energy a strong $\text{S}3p$ density of states is found. Above the Fermi level we also find a broad resonance with a large weight on the $\text{S}3p$ orbitals, although strongly hybridized with Ru states in an anti-bonding S–Ru interaction. The resonance maximum lies ~ 2 eV above the Fermi level, which is marginally higher than the experimentally observed resonance maximum at 1.68 ± 0.1 eV above the Fermi level.

Calculation of the charge transfer times. The charge transfer dynamics are computed using the electronic hamiltonian obtained in the density functional calculations previously described. The initial electron wave packet $|\phi_R\rangle$ is constructed as a projection of a linear combination of the $\text{S}3p$ orbitals onto the unoccupied bands at the energies of the resonance region. The wavefunction of the resonance depends on the excitation process. Since the electron is excited from a state of s -symmetry, the admixture of p_x , p_y and p_z components is given by the direction of the electric field vector of the incoming radiation $\mathbf{E}$. Therefore we take $|\phi_R\rangle = |\phi(t=0)\rangle \propto \sum_i E_i |p_i\rangle$. From the time evolution of the wave packet we can calculate the probability of finding the electron in the initial state $P(t) = |A(t)|^2$, where $A(t) = \langle \phi_R | \phi(t) \rangle$ is the so-called survival amplitude. The Fourier transform $A(t)$ is directly related to the projection of the Green function onto the initial state $\tilde{A}(\omega) \propto \langle \phi_R | \hat{G}(\omega) | \phi_R \rangle = G_{RR}(\omega)$ (ref. 26). The characteristic decay time of the resonance population τ is then computed using two procedures. Either the width of the resonance Δ is directly estimated from $G_{RR}(\omega)$ and $\tau = \hbar\Delta^{-1}$, or $P(t)$ is transformed into real time and τ is defined such that $P(t) \leq 1/e$ if $t \geq \tau$. Both methods produce very similar results. However, we prefer the second method since τ is obtained directly and it is not necessary to assume the lorentzian line-shape. We should point out that $G_{RR}(\omega)$ has to be calculated for the semi-infinite system, that is, we have to get rid of the finite size effects associated with the slab calculations. This is instrumental in getting reliable resonance widths and lifetimes. This is done by combining the information from a slab calculation with the *ab initio* hamiltonian obtained for bulk Ru and using recursive techniques to calculate $G_{RR}(\omega)$. We have assigned an error bar of 0.15 fs to our theoretical values. This reflects both the presence of two

non-equivalent sulphur atoms in the surface and the numerical accuracy of our procedure.

Received 18 January; accepted 20 May 2005.

1. Zewail, A. H. Femtochemistry: atomic-scale dynamics of the chemical bond (adapted from the Nobel lecture). *J. Phys. Chem. A* **104**, 5660–5694 (2000).
2. Hentschel, M. *et al.* Attosecond metrology. *Nature* **414**, 509–513 (2001).
3. Drescher, M. *et al.* Time-resolved atomic inner-shell spectroscopy. *Nature* **419**, 803–807 (2002).
4. Baltuska, A. *et al.* Attosecond control of electronic processes by intense light fields. *Nature* **421**, 611–625 (2003).
5. Steinmeyer, G., Sutter, D. H., Gallmann, L., Matuschek, N. & Keller, U. Frontiers in ultrashort pulse generation: Pushing the limits in linear and nonlinear optics. *Science* **286**, 1507–1512 (1999).
6. Petek, H. & Ogawa, S. Femtosecond time-resolved two-photon photoemission studies of electron dynamics in metals. *Prog. Surf. Sci.* **56**, 239–310 (1997).
7. Kienberger, R. *et al.* Atomic transient recorder. *Nature* **427**, 817–821 (2004).
8. Goulielmakis, E. *et al.* Direct measurement of light waves. *Science* **305**, 1267–1269 (2004).
9. Björneholm, O., Nilsson, A., Sandell, A., Hernnäs, B. & Martensson, N. Determination of time scales for charge-transfer screening in physisorbed molecules. *Phys. Rev. Lett.* **68**, 1892–1895 (1992).
10. Ohno, M. Deexcitation processes in adsorbates. *Phys. Rev. B* **50**, 2566–2575 (1994).
11. Björneholm, O. *et al.* Femtosecond dissociation of core-excited HCl monitored by frequency detuning. *Phys. Rev. Lett.* **79**, 3150–3153 (1997).
12. Keller, C. *et al.* Ultrafast charge transfer times of chemisorbed species from Auger resonant Raman studies. *Phys. Rev. Lett.* **80**, 1774–1777 (1998).
13. Keller, C. *et al.* Femtosecond dynamics of adsorbate charge-transfer processes as probed by high-resolution core-level spectroscopy. *Phys. Rev. B* **57**, 11951–11954 (1998).
14. Feifel, R. *et al.* Observation of a continuum-continuum interference hole in ultrafast dissociating core-excited molecules. *Phys. Rev. Lett.* **85**, 3133–3136 (2000).
15. Wurth, W. & Menzel, D. Ultrafast electron dynamics at surfaces probed by resonant Auger spectroscopy. *Chem. Phys.* **251**, 141–149 (2000).
16. Brühwiler, P. A., Karis, O. & Mårtensson, N. Charge-transfer dynamics studied using resonant core spectroscopies. *Rev. Mod. Phys.* **74**, 703–740 (2002).
17. Schnadt, J. *et al.* Experimental evidence for sub-3-fs charge transfer from an aromatic adsorbate to a semiconductor. *Nature* **418**, 620–623 (2002).
18. Föhlisch, A. *et al.* Energy dependence of resonant charge transfer from adsorbates to metal substrates. *Chem. Phys.* **289**, 107–115 (2003).
19. Keller, C. *et al.* Electronic transfer processes studied at different time scales by selective resonant core hole excitation of adsorbed molecules. *Appl. Phys. A* **78**, 125–129 (2004).
20. Coville, M. & Thomas, T. D. Molecular effects on inner-shell lifetimes: Possible test of the one-center model of Auger decay. *Phys. Rev. A* **43**, 6053–6056 (1991).
21. Schwennicke, C., Jürgens, D., Held, G. & Pfnür, H. The structure of dense sulphur layers on Ru(0001) I. The $c(2 \times 4)$ structure. *Surf. Sci.* **316**, 81–91 (1994).
22. Jürgens, D., Schwennicke, C. & Pfnür, H. Surface structure analysis of the domain-wall phase of S/Ru(0001) using an efficient parameter optimization method. *Surf. Sci.* **381**, 174–189 (1997).
23. Krause, M. O. & Oliver, J. H. Natural widths of atomic K and L levels, K alpha X-ray lines and several KLL auger lines. *J. Phys. Chem. Ref. Data* **8**, 329–338 (1979).
24. Sánchez-Portal, D., Artacho, E., Ordejón, P. & Soler, J. M. Density-functional method for very large systems with LCAO basis sets. *Int. J. Quant. Chem.* **65**, 453–461 (1997).
25. Soler, J. M. *et al.* The SIESTA method for *ab initio* order-N materials simulation. *J. Phys. Condens. Matter* **14**, 2745–2779 (2002).
26. Borisov, A. G., Kazansky, A. K. & Gaquay, J. P. Resonant charge transfer in ion-metal surface collisions: Effect of a projected band gap in the H-Cu(111) system. *Phys. Rev. B* **59**, 10935–10949 (1999).

Acknowledgements We acknowledge support by the staff of MAX-lab, Lund, Sweden, in particular J. N. Andersen and the ARI program. This work was supported by the Deutsche Forschungsgemeinschaft under Schwerpunktprogramm 1093 “Dynamik von Elektronentransferprozessen an Grenzflächen”, the Basque Departamento de Educación, the University of the Basque Country, the Spanish MEC, European Network of Excellence NANOQUANTA, and Max-Planck Awards for Scientific Cooperation to P.M.E. and D.M.

Author Information Reprints and permissions information is available at ngp.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.W. (wilfried.wurth@desy.de).

Spin transition of iron in magnesiowüstite in the Earth's lower mantle

Jung-Fu Lin^{1,†}, Viktor V. Struzhkin¹, Steven D. Jacobsen¹, Michael Y. Hu², Paul Chow², Jennifer Kung³, Haozhe Liu², Ho-kwang Mao¹ & Russell J. Hemley¹

Iron is the most abundant transition-metal element in the mantle and therefore plays an important role in the geochemistry and geodynamics of the Earth's interior^{1–11}. Pressure-induced electronic spin transitions of iron occur in magnesiowüstite, silicate perovskite and post-perovskite^{1–4,8,10,11}. Here we have studied the spin states of iron in magnesiowüstite and the isolated effects of the electronic transitions on the elasticity of magnesiowüstite with *in situ* X-ray emission spectroscopy and X-ray diffraction to pressures of the lowermost mantle. An observed high-spin to low-spin transition of iron in magnesiowüstite results in an abnormal compressional behaviour between the high-spin and the low-spin states. The high-pressure, low-spin state exhibits a much higher bulk modulus and bulk sound velocity than the low-pressure, high-spin state; the bulk modulus jumps by ~35 per cent and bulk sound velocity increases by ~15 per cent across the transition in (Mg_{0.83},Fe_{0.17})O. Although no significant density change is observed across the electronic transition, the jump in the sound velocities and the bulk modulus across the transition provides an additional explanation for the seismic wave heterogeneity in the lowermost mantle^{12–21}. The transition also affects current interpretations of the geophysical and geochemical models using extrapolated or calculated thermal equation-of-state data without considering the effects of the electronic transition^{5,6,22,23}.

The electronic spin transition of iron in magnesiowüstite and silicate perovskite in the Earth's lower mantle has been postulated to have major geophysical and geochemical consequences: causing a large density change, shifting the partitioning of iron between magnesiowüstite and perovskite, altering radiative thermal conductivity, and enhancing compositional layering in the lower mantle^{1–4,8,10,11}. However, no significant displacement in iron partitioning between magnesiowüstite and perovskite has been observed experimentally^{5–7}, and the effects of the electronic transition on the density and elasticity have not previously been measured. Seismic observations show heterogeneities in seismic-wave velocities in the Earth's lower mantle^{12–18}, where compositional and thermal variation, partial melting, and a phase transition in perovskite have been used to explain the origin of the heterogeneity^{16–21}. Assuming a relatively large volume change across the spin transition and neglecting the spin-pairing energy of ~4 eV in thermodynamic calculations, it has been suggested that the partition coefficient of iron between magnesiowüstite and perovskite increases by several orders of magnitude in the mid- to lower mantle⁸, although *in-situ* X-ray diffraction and quenched sample analyses have not demonstrated such a dramatic change under mid- to lower-mantle conditions^{5–7}.

To understand the consequences of the electronic transition on the geophysical and geochemical processes of the Earth's interior, we

have studied the spin states of iron in magnesiowüstite and the effects of the electronic transitions on the elasticity of magnesiowüstite with *in situ* X-ray emission spectroscopy (XES) and X-ray diffraction under lower-mantle pressures.

The spin states of iron in magnesiowüstite with varying Fe-content were probed by XES in a diamond anvil cell (Fig. 1). XES is an established technique that provides direct information on the local magnetic properties of iron atoms and has been widely used to study magnetic transitions in iron-containing systems^{8,10,11}. The magnetic state of Fe is characterized by the appearance of the satellite emission peak ($K\beta'$) located at the lower-energy region of the main emission

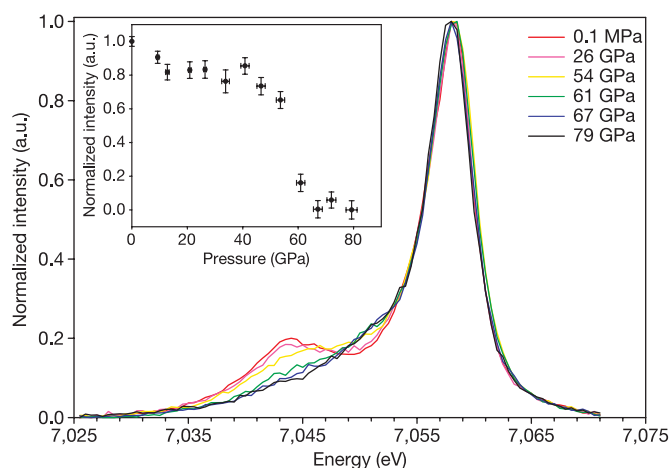


Figure 1 | Representative X-ray emission spectra of Fe- $K\beta$ collected from a single-crystal magnesiowüstite, (Mg_{0.75},Fe_{0.25})O, in (110) orientation at high pressures. The spectrum at ambient conditions was measured outside the diamond anvil cell. The spectra were normalized to unity and shifted in energy to compensate for the pressure-induced shift of the line maximum, based on the main fluorescence peak ($K\beta$) at 7,058 eV. The presence of the satellite peak ($K\beta'$) below 55 GPa is characteristic of the magnetic state of iron, whereas the absence of the satellite peak above 67 GPa indicates the collapse of the magnetization. Inset, normalized intensity of the satellite peak of iron in magnesiowüstite as a function of pressure. The intensity of the satellite peak was obtained by subtracting each spectrum from the one at the highest pressure (low-spin state). The errors in integrated intensity were propagated from statistical errors in original spectra (errors span two standard deviations). The intensity of the satellite peak is proportional to the magnetic moment of the iron atoms^{8,10,11}, so the change in the intensity of the satellite peak can be used to understand the electronic spin transition of iron in magnesiowüstite. The observed electronic spin transition pressure for (Mg_{0.75},Fe_{0.25})O is close to the transition pressure in (Mg_{0.83},Fe_{0.17})O (ref. 8).

¹Geophysical Laboratory, Carnegie Institution of Washington, 5251 Broad Branch Road NW, Washington DC 20015, USA. ²HPCAT, Carnegie Institution of Washington, Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439, USA. ³The Mineral Physics Institute, University of New York at Stony Brook, Stony Brook, New York 11794, USA. [†]Present address: Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, California 94550, USA.

peak ($K\beta_{1,3}$) of $\sim 7,058$ eV, which results from the $3p$ – $3d$ core–hole exchange interaction in the final state of the emission process. On the other hand, the collapse of the magnetization of Fe is characterized by the disappearance of the low-energy satellite owing to the loss of the $3d$ magnetic moment. XES measurements of the Fe $K\beta$ line were carried out at the High Pressure Collaborative Access Team (HPCAT) sector of the Advanced Photon Source (APS) at the Argonne National Laboratory. A monochromatic X-ray beam of 12 keV was focused down to $20\ \mu\text{m}$ vertically and $60\ \mu\text{m}$ horizontally at the sample position. The Fe $K\beta$ emission spectra were collected through a Be gasket by an one-metre Rowland circle spectrometer in the vertical scattering geometry. A Si (333) single-crystal wafer glued onto a spherical substrate of one-metre radius was used as the analyser and a Peltier-cooled silicon detector (Amptek XR 100CR) was used to detect the emitted X-ray fluorescence. The sample was sandwiched between two layers of NaCl, and pressures were measured from rubies placed in the NaCl medium using the ruby pressure scale. The details of the experimental set-up^{8,10,11} and sample syntheses²⁴ are reported elsewhere.

The XES spectra of the Fe $K\beta$ fluorescence lines in $(\text{Mg}_{0.75}\text{Fe}_{0.25})\text{O}$ and $(\text{Mg}_{0.40}\text{Fe}_{0.60})\text{O}$ reveal that a high-spin to low-spin transition occurs at 54 to 67 GPa and 84 to 102 GPa, respectively (Fig. 1) (see

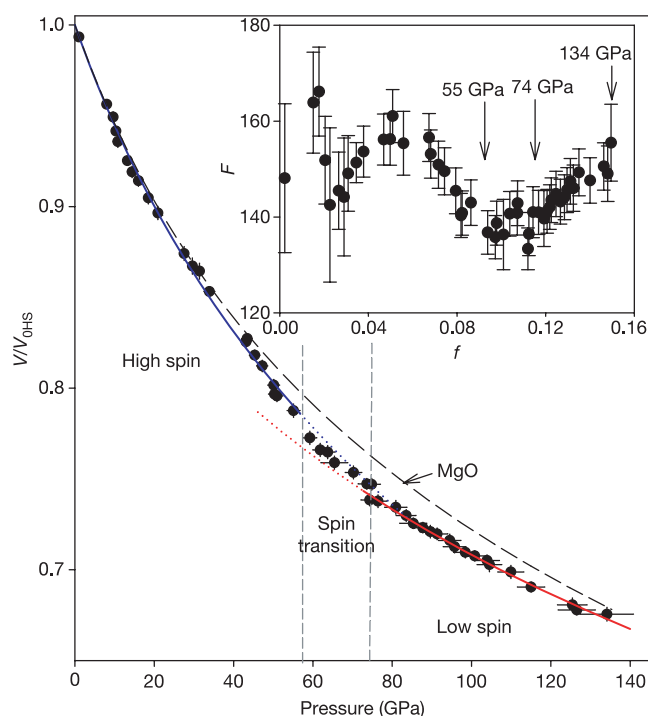


Figure 2 | Normalized volume of magnesiowüstite, $(\text{Mg}_{0.83}\text{Fe}_{0.17})\text{O}$, as a function of pressure at 300 K. A weighted least-squares fit to the high-spin state and low-spin state, based on the BM EOS (ref. 27), show that the K_{0T} and K_{0T}' of the high-spin state are $160.7(\pm 3.7)$ and $3.28(\pm 0.21)$ GPa (blue curve), respectively, whereas the K_{0T} of the low-spin state is $250(\pm 28)$ GPa with a V_{0LS}/V_{0HS} of $0.904(\pm 0.016)$ and an assumed K_{0T}' of 4 (red curve). V_{0HS} and V_{0LS} are the volumes of the high-spin state and low-spin state at ambient conditions, respectively. Based on the Vinet EOS (ref. 28), the K_{0T} and K_{0T}' of the high-spin state are $161.3(\pm 5.2)$ and $3.25(\pm 0.37)$ GPa, respectively, whereas the K_{0T} of the low-spin state is $245(\pm 21)$ GPa, assuming a K_{0T}' of 4. The dashed black line represents the EOS of MgO with a K_{0T} of 160 GPa and K_{0T}' of 4 for comparison (ref. 25). The sample remains in the NaCl-type structure at all pressures⁹, and no significant change in density is observed. Inset, normalized stress (F) versus normalized strain (f) plot²⁷. F decreases with increasing f up to 55 GPa, whereas F increases with increasing f above 74 GPa, indicating a significant change in the incompressibility from the high-spin state to the low-spin state. Error bars span two standard deviations.

Supplementary Information for details). To understand the isolated effect of the spin transition of iron on the elasticity of magnesiowüstite, we also carried out *in situ* X-ray powder diffraction experiments in a diamond anvil cell (Fig. 2). A polycrystalline magnesiowüstite sample—either $(\text{Mg}_{0.83}\text{Fe}_{0.17})\text{O}$ or $(\text{Mg}_{0.40}\text{Fe}_{0.60})\text{O}$ —was loaded into a diamond anvil cell with a neon pressure medium and Pt pressure scale^{25,26}. The iron concentrations in $(\text{Mg}_{0.75}\text{Fe}_{0.25})\text{O}$ and $(\text{Mg}_{0.83}\text{Fe}_{0.17})\text{O}$ samples are close to the iron concentration in magnesiowüstite in the lower mantle, whereas the $(\text{Mg}_{0.40}\text{Fe}_{0.60})\text{O}$ sample is used to understand the compositional effect on the spin transition and volume change. A focused monochromatic beam ($0.4246\ \text{\AA}$) with a beam diameter of approximately $10\ \mu\text{m}$ was used as the X-ray source for angle-dispersive X-ray diffraction experiments, and the diffracted X-ray was collected with an image plate (MAR345). Three diffraction peaks of (111), (200), and (220) were used to calculate the cell parameters of magnesiowüstite, and pressures were calculated from the equation of state (EOS) of Pt (refs 26, 27). Because the ruby pressure scale used during the XES studies was originally calibrated against the Pt pressure scale²⁶, use of the Pt scale during X-ray diffraction measurements should yield precisely the same pressure as that obtained during the XES measurements. The pressure and volume data were analysed with the Birch–Murnaghan (BM) EOS and the Vinet EOS (refs 27, 28) using a weighted least-squares linear fit—the eulerian strain (f) and the normalized stress (F) in the BM EOS and $\ln(H)$ and $1 - X$ in the Vinet EOS—to obtain values for the isothermal bulk modulus (K_T) at ambient conditions (K_{0T}) and pressure derivative of the bulk modulus at ambient conditions (K_{0T}') (Fig. 2).

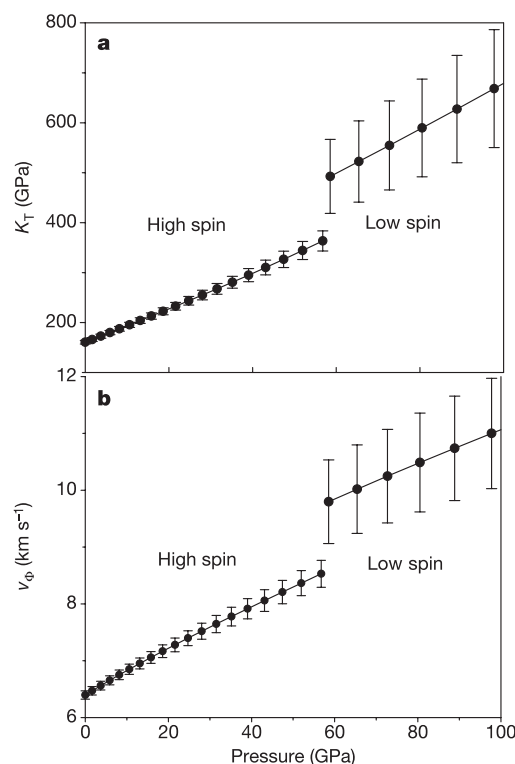


Figure 3 | Calculated isothermal bulk modulus and bulk sound velocity as a function of pressure for the high-spin and low-spin states. **a**, Bulk modulus; **b**, bulk sound velocity. Although no significant volume/density change was observed at approximately 60 GPa (ref. 8), K_T jumps by $\sim 35\%$ across the transition. The bulk sound velocity ($v_b = \sqrt{K_S/\rho}$; $K_S = K_T(1 + \alpha\gamma T)$; α is the thermal expansion coefficient and γ is the Grüneisen parameter) increases by $\sim 15\%$ across the transition. These values and their errors are calculated based on the BM EOS. These calculations based on the Vinet EOS show similar results. Error bars span two standard deviations.

Within experimental uncertainties, no significant volume or density change was observed in $(\text{Mg}_{0.83}\text{Fe}_{0.17})\text{O}$ at approximately 60 GPa where the high-spin to low-spin transition has been observed by ref. 8 using XES (Figs 1, 2). On the other hand, a volume decrease of $\sim 1.6\%$ was observed in $(\text{Mg}_{0.40}\text{Fe}_{0.60})\text{O}$ at 95 GPa, consistent with the high-spin to low-spin transition pressure observed in XES studies. Our results show that addition of FeO in MgO stabilizes the high-spin state to higher pressures. The difference in effective ionic radii between octahedrally coordinated high-spin and low-spin Fe^{2+} in sulphides (based on the effective ionic radius of O^{2-} in six-fold coordination of 1.40 Å) is 0.16 Å, resulting in a volume change of 21% at ambient pressure²⁹. It has been suggested that the electronic spin transition of Fe in iron-containing compounds can cause a significant volume reduction across the transition^{1,2,8,29}. In Fe_2O_3 , the volume change due to the electronic spin transition is estimated to be $\sim 7\%$; however, a concurrent structure transition in Fe_2O_3 complicates the estimation of the volume reduction³⁰. It is believed that the composition of magnesiowüstite in the lower mantle^{5–7,22,23} contains approximately 20% FeO, which corresponds to a density change of only $\sim 0.5\%$ across the spin transition, according to our results. Considering that the lower mantle contains only approximately 20% of magnesiowüstite and that high temperature would further reduce the electronic effect on the density reduction, the very small density change is likely to be under the detection limit of current seismic-wave resolution.

By plotting the $(\text{Mg,Fe})\text{O}$ compression data as normalized stress (F) against eulerian strain (f), it is possible to see the change in compressibility across the electronic transition (Fig. 2). These analyses show that the electronic transition significantly affects the incompressibility of magnesiowüstite at lower-mantle pressures. As shown, the normalized stress decreases with increasing the eulerian strain up to 55 GPa, whereas the normalized stress increases with increasing the eulerian strain above 74 GPa. We note that an electronic spin transition is observed in $(\text{Mg}_{0.83}\text{Fe}_{0.17})\text{O}$ between 60 to 70 GPa (ref. 8). A weighted least-squares fit to the high-spin state and low-spin state, based on the BM EOS, shows that the K_{OT} and K_{OT}' of the high-spin state are $160.7(\pm 3.7)$ and $3.28(\pm 0.21)$ GPa, respectively, whereas the K_{OT} of the low-spin state is $250(\pm 28)$ GPa with a $V_{\text{OLS}}/V_{\text{OHS}}$ of $0.904(\pm 0.016)$ and an assumed K_{OT}' of 4, where V_{OHS} and V_{OLS} are the volumes of the high-spin state and low-spin state at ambient conditions, respectively. These parameters remain similar on the basis of the Vinet EOS analyses (ref. 28), indicating that the incompressibility change does not depend on the form of the EOS used. The calculated K_{T} and bulk sound velocity ($v_{\Phi} = \sqrt{(K_{\text{S}}/\rho)}$, where K_{S} is the adiabatic bulk modulus and ρ is the density) show that a significant change in K_{T} and v_{Φ} occurs across the transition and that the low-spin state has higher K_{T} and v_{Φ} than the high-spin state (Fig. 3); K_{T} increases by $\sim 35\%$ and v_{Φ} increases by $\sim 15\%$ across the transition.

The volume difference, change of entropy, and spin-pairing energy across the electronic transition are key parameters in estimating the thermodynamics of the spin transition, in particular Clapeyron slope and the partitioning of iron^{1,8,31,32}. On the basis of our current XES and X-ray diffraction results (a small volume change) and thermodynamic considerations (the spin-pairing energy of 4.1 eV is considered in this study)^{31,32}, we found that, for a magnesiowüstite containing $\sim 20\%$ FeO, the Clapeyron slope of the high-spin to low-spin transition is approximately 56 K per GPa and the spin transition is shifted to ~ 120 GPa at 2,500 K (close to the top of the D'' region) (see Supplementary Information for details)³².

The unusual effects of the spin transition in iron on the elastic properties of magnesiowüstite reported here have several implications for the seismology, dynamics and geochemistry of the lower mantle. Recent seismic observations have shown that seismic wave heterogeneities exist in the Earth's lowermost mantle^{15,16,18}; in particular, the shear-wave velocity (v_{S}) changes more significantly than K_{S} , v_{P} and v_{Φ} . The changes in v_{Φ} and v_{S} are anti-correlated¹².

The phase transition from perovskite to post-perovskite has been used to explain the seismic-wave discontinuity in the D'' layer^{19–21}. Our results show that the electronic transition of iron in magnesiowüstite is likely to occur in the lowermost mantle and that the K_{S} and v_{Φ} in magnesiowüstite would increase significantly across the transition in the lowermost mantle. Although the relatively less amount of magnesiowüstite in the lower mantle (approximately 20%) would reduce the magnitude of the effect on the elasticity, the electronic transition is still likely to leave a signature in the change of the sound velocities. Such a change in sound velocities provides an additional explanation for the seismic wave heterogeneities in the lowermost mantle^{15,16,18}. The jump in v_{Φ} across the electronic transition implies changes in v_{P} and v_{S} , so future ultrasonic or Brillouin studies on the v_{P} and v_{S} across the transition would provide crucial information in understanding the contribution of the electronic transition to this seismic-wave heterogeneity.

The electronic spin transition in iron has also been reported for both ferromagnesian silicate perovskite and the post-perovskite phase¹⁰. The electronic transitions should also affect the incompressibility and sound velocities in perovskite and post-perovskite. We also observed optically that the opacity of the magnesiowüstite increases with increasing pressure and the low-spin state can be well heated by the radiation of the Nd:YLF laser at 1,053 nm; the heated magnesiowüstite sample across the transition remains in the NaCl-type structure, as revealed from *in situ* X-ray diffraction. The enhanced opacity of $(\text{Mg,Fe})\text{O}$ above the transition is at odds with that predicted theoretically². One possible explanation is that in ref. 2 only the average energy of the excited singlet states of low-spin Fe^{2+} was estimated. Because the post-perovskite is suggested to be transparent in the infrared region, the low-spin state of magnesiowüstite may become dominant in blocking radiative heat transfer in the lowermost mantle.

Our current understanding of the EOS for iron-bearing minerals comprising the bulk of the Earth's lower mantle, namely magnesiowüstite and ferromagnesian silicate perovskite, appears to be inadequate, considering the effects of the pressure-induced spin-state transitions in Fe. We present experimental evidence for the anomalous compressibility of magnesiowüstite through the high-spin to low-spin transition, indicating that an extrapolation of the low-pressure, high-spin EOS inaccurately predicts the density of magnesiowüstite in deeper parts of the lower mantle. Furthermore, a discontinuous jump in the K_{T} and v_{Φ} values measured in this study may account for the seismic-wave heterogeneities in the lowermost mantle. It remains to be seen how the spin-state transitions in the lower mantle will affect shear velocities of these phases, but the emerging picture of the Earth's deeper mantle from complementary seismic and mineral physics studies suggest that it is a more complex region than traditionally thought.

Received 17 January; accepted 13 May 2005.

1. Sherman, D. M. in *Structural and Magnetic Phase Transitions in Minerals* (eds Ghose, S., Coey, J. M. D. & Salje, E.) 113–118 (Springer, New York, 1988).
2. Sherman, D. M. The high-pressure electronic structure of magnesiowüstite (Mg,FeO): applications to the physics and chemistry of the lower mantle. *J. Geophys. Res.* **96** (B9), 14299–14312 (1991).
3. Sherman, D. M. & Jansen, H. J. F. First-principles predictions of the high-pressure phase transition and electronic structure of FeO: implications for the chemistry of the lower mantle and core. *Geophys. Res. Lett.* **22**, 1001–1004 (1995).
4. Cohen, R. E., Mazin, I. I. & Isaak, D. G. Magnetic collapse in transition metal oxides at high pressure: implications for the Earth. *Science* **275**, 654–657 (1997).
5. Mao, H. K., Shen, G. & Hemley, R. J. Multivariable dependence of Fe-Mg partitioning in the lower mantle. *Science* **278**, 2098–2100 (1997).
6. Andraut, D. Evaluation of (Mg,Fe) partitioning between silicate perovskite and magnesiowüstite up to 120 GPa and 2300 K. *J. Geophys. Res.* **106**, 2079–2087 (2001).
7. Kesson, S. E., Fitz Gerald, J. D., O'Neill, H., St. C. & Shelley, J. M. G. Partitioning of iron between magnesian silicate perovskite and magnesiowüstite at about 1 Mbar. *Phys. Earth Planet. Inter.* **131**, 295–310 (2002).

8. Badro, J. *et al.* Iron partitioning in Earth's mantle: toward a deep lower mantle discontinuity. *Science* **300**, 789–791 (2003).
9. Lin, J. F. *et al.* Stability of magnesiowüstite in the Earth's lower mantle. *Proc. Natl Acad. Sci. USA* **100**, 4405–4408 (2003).
10. Badro, J. *et al.* Electronic transitions in perovskite: possible nonconvecting layers in the lower mantle. *Science* **305**, 383–386 (2004).
11. Li, J. *et al.* Electronic spin state of iron in lower mantle perovskite. *Proc. Natl Acad. Sci. USA* **101**, 14027–14030 (2004).
12. Su, W. J. & Dziewonski, A. M. Simultaneous inversion for 3-D variations in shear and bulk velocity in the mantle. *Phys. Earth Planet. Inter.* **100**, 135–156 (1997).
13. Kellogg, L. H., Hager, B. H. & van der Hilst, R. D. Compositional stratification in the deep mantle. *Science* **283**, 1881–1884 (1999).
14. van der Hilst, R. D. & Kárason, H. Compositional heterogeneity in the bottom 1000 kilometers of Earth's mantle: toward a hybrid convection model. *Science* **283**, 1885–1888 (1999).
15. Garnero, E. Heterogeneity of the lowermost mantle. *Annu. Rev. Earth Planet. Sci.* **28**, 509–537 (2000).
16. Masters, G., Laske, G., Bolton, H. & Dziewonski, A. M. in *Earth's Deep Interior: Mineral Physics and Tomography From the Atomic to the Global Scale* (eds Karato, S., Forte, A. M., Liebermann, R. C., Masters, G. & Stixrude, L.) 63–87 (American Geophysical Union, Washington DC, 2000).
17. Karato, S. I. & Kaiki, B. B. Origin of lateral variation of seismic wave velocities and density in the deep mantle. *J. Geophys. Res.* **106**, 21771–21783 (2001).
18. Lay, T., Garnero, E. J. & Williams, Q. Partial melting in a thermo-chemical boundary layer at the base of the mantle. *Phys. Earth Planet. Inter.* **146**, 441–467 (2004).
19. Murakami, M., Hirose, K., Kawamura, K., Sata, N. & Ohishi, Y. Post-perovskite phase transition in MgSiO₃. *Science* **304**, 855–858 (2004).
20. Oganov, A. R. & Ono, S. Theoretical and experimental evidence for a post-perovskite phase of MgSiO₃ in Earth's D'' layer. *Nature* **430**, 445–448 (2004).
21. Tsuchiya, T., Tsuchiya, J., Umemoto, K. & Wentzcovitch, R. M. Elasticity of post-perovskite MgSiO₃. *Geophys. Res. Lett.* **31**, L14603, doi:10.1029/2004GL020278 (2004).
22. Jackson, I. Elasticity, composition and temperature of the Earth's lower mantle: a reappraisal. *Geophys. J. Int.* **134**, 291–311 (1998).
23. Stacey, F. D. & Isaak, D. G. Compositional constraints on the equation of state and thermal properties of the lower mantle. *Geophys. J. Int.* **146**, 143–154 (2001).
24. Jacobsen, S. D. *et al.* Structure and elasticity of single-crystal (Mg,Fe)O and a new method of generating shear waves for gigahertz ultrasonic interferometry. *J. Geophys. Res.* **107**(B2), 10.1029/2001JB000490 (2002).
25. Speziale, S., Zha, C. S., Duffy, T. S., Hemley, R. J. & Mao, H. K. Quasi-hydrostatic compression of magnesium oxide to 52 GPa: Implications for the pressure-volume-temperature equation of state. *J. Geophys. Res.* **106**, 515–528 (2001).
26. Holmes, N. C., Moriarty, J. A., Gathers, G. R. & Nellis, W. J. The equation of state of platinum to 660 GPa (6.6 Mbar). *J. Appl. Phys.* **66**, 2962–2967 (1989).
27. Birch, F. Equation of state and thermodynamic parameters of NaCl to 300 kbar in the high-temperature domain. *J. Geophys. Res.* **91**, 4949–4954 (1986).
28. Vinet, P., Ferrante, J., Rose, J. H. & Smith, J. R. Compressibility of solids. *J. Geophys. Res.* **92**, 9319–9325 (1987).
29. Shannon, R. D. & Prewitt, C. T. Effective ionic radii in oxides and fluorides. *Acta Crystallogr.* **B25**, 925–946 (1969).
30. Badro, J. *et al.* Nature of the high-pressure transition in Fe₂O₃ hematite. *Phys. Rev. Lett.* **89**, 205504 (2002).
31. Burns, R. G. *Mineralogical Applications of Crystal Field Theory* Ch. 2, 7–43 (Cambridge Univ. Press, Cambridge, 1993).
32. Brown, J. M. & Shankland, T. J. Thermodynamic parameters in the Earth as determined from seismic profiles. *Geophys. J. R. Astron. Soc.* **66**, 579–596 (1981).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Caracas, R. Cohen, G. Shen, V. Prakapenka, W. Sturhahn, J. M. Jackson, P. Silver, B. Militzer, S. Hardy, C. Prewitt, M. Somayazulu, P. Dera, and Y. Fei for discussions, S. J. Mackwell for help with sample synthesis, HPCAT for the use of the X-ray facilities, and GSECARS, APS, for the use of the Raman system. This work and use of the APS are supported by the US Department of Energy, Basic Energy Sciences, Office of Science and the State of Illinois under HECA. Work at Carnegie was supported by DOE/BES, DOE/NNSA (CDAC), NSF, and the W. M. Keck Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.-F.L. (j.lin@gl.ciw.edu).

The long-term strength of Europe and its implications for plate-forming processes

M. Pérez-Gussinyé¹† & A. B. Watts¹

Field-based geological studies show that continental deformation preferentially occurs in young tectonic provinces rather than in old cratons¹. This partitioning of deformation suggests that the cratons are stronger than surrounding younger Phanerozoic provinces. However, although Archaean and Phanerozoic lithosphere differ in their thickness^{2–4} and composition^{4,5}, their relative strength is a matter of much debate. One proxy of strength is the effective elastic thickness of the lithosphere, T_e . Unfortunately, spatial variations in T_e are not well understood, as different methods yield different results. The differences are most apparent in cratons, where the ‘Bouguer coherence’ method yields large T_e values (>60 km)^{6–9} whereas the ‘free-air admittance’ method yields low values (<25 km)¹⁰. Here we present estimates of the variability of T_e in Europe using both methods. We show that when they are consistently formulated¹¹, both methods yield comparable T_e values that correlate with geology, and that the strength of old lithosphere (≥ 1.5 Gyr old) is much larger (mean $T_e > 60$ km) than that of younger lithosphere (mean $T_e < 30$ km). We propose that this strength difference reflects changes in lithospheric plate structure (thickness, geothermal gradient and composition) that result from mantle temperature and volatile content decrease through Earth’s history.

The current debate concerning the strength of continental lithosphere is focused on the values of T_e estimated using both forward and inverse (spectral) methods based on the Bouguer coherence and free-air admittance. For example, forward models in the Slave craton (Canadian shield) reveal a T_e of about 12 km (ref. 12), which is much smaller than the ~ 100 km obtained using Bouguer coherence in the same area⁶. This strength difference arises because, in cratons, forward models and Bouguer coherence yield estimates of the strength at different times¹². Forward models are generally based on the reconstruction of the original (surface and subsurface) loads and their associated flexures and reveal the strength at the time of a specific loading event (for example, orogeny or rifting), while Bouguer coherence is based on present-day topography and gravity anomaly data and yields the current strength of thick, cooled, cratonic lithosphere¹². Because the free-air admittance is also based on present-day topography and gravity anomaly data, it should yield a T_e similar to that obtained using the Bouguer coherence. However, in some regions, it yields $T_e < 25$ km (ref. 10), while the Bouguer coherence yields values in excess of 60 km (refs 6–9).

Recently, it has been shown that this discrepancy occurs because the free-air admittance and Bouguer coherence methods have not been consistently formulated¹¹. The calculation of both functions involves the estimation of the spectra of finite and non-periodic data. Therefore, the wavelength dependence of the admittance and coherence varies with the size of the data window analysed. In the admittance method, analytical solutions of the predicted spectra (which correspond to infinite data windows) have been compared to

observed spectra of the data within a finite window¹¹. When the underlying T_e is large (>30 – 40 km), this can lead to an underestimation of T_e by a factor of 2 (ref. 11). However, when the observed and predicted admittance functions are calculated in the same data windows, as is usually the case in the coherence method, then the results from the two techniques are equivalent¹¹.

We present here estimates of the T_e structure of Europe using both Bouguer coherence and free-air admittance. The resulting T_e structures are similar, and correlate well with the tectonic provinces in Europe inferred from geological and other geophysical data¹³ (Fig. 1). High- T_e regions correlate with Precambrian Baltica and Avalonia. Low- T_e regions, in contrast, correlate with the younger provinces accreted during the Caledonian, Variscan and Alpine orogenies. These results are in agreement with local studies of T_e in Europe^{14–17}. Figure 1 also shows that the largest changes in T_e occur at the sutures that separate different provinces. Deep seismic data indicate that these sutures coincide with major tectonic boundaries in the crust and shallow lithospheric mantle (for example, Iapetus suture¹⁸, Thor suture¹⁹). Our results show that, in addition, these sutures correlate with major changes in lithospheric strength.

The recovered T_e is generally consistent with other physical properties of the lithosphere. High- T_e regions generally correlate with areas of large thermal thickness (as derived from heat flow data³) and fast seismic S-wave velocities²⁰ and vice versa (Fig. 1c, d). This indicates that T_e is high in regions where the lithosphere is thick and cold. In addition, a close correlation exists between T_e and seismicity: high- and low- T_e regions show sparse and abundant seismicity, respectively (Fig. 1d). This suggests that in high- T_e areas, the strength of the lithosphere is large enough to reduce the background level of seismicity²¹.

The dependence of T_e on age is shown in Fig. 2. The figure shows a plot of the mean T_e obtained from Bouguer coherence for each of the main tectonic provinces versus their age (Table 1). Unlike T_e values in oceanic lithosphere, spectrally derived continental T_e estimates do not necessarily reflect the strength at the time of loading. This is because spectral estimates reflect the present-day response to loads, which resulted from specific geological events (for example, orogeny and rifting) as well as from subsequent sedimentation and erosion. The relatively young Caledonian, Variscan and Alpine provinces still have significant topography such that their spectrally derived T_e also mainly reflects the strength of the basement at the time of these orogenies (that is, it is the T_e at the time of loading which is recovered). However, cratons have subdued topography, and so their spectrally derived T_e mainly reflects the response to post-orogenic erosion and sedimentation, both of which may continue up to the present day. Because forward models reflect the T_e at the time of loading¹², they generally coincide with estimates of spectrally derived T_e in young tectonic provinces (see, for example, ref. 17) but differ in cratonic areas¹².

¹Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3PR, UK. †Present address: Institut de Ciències de la Terra ‘Jaume Almera’, Lluís Solé i Sabarís, s/n, 08028 Barcelona, Spain.

Figure 2 shows that Archaean and Early/Middle Proterozoic tectonic provinces (≥ 1.5 Gyr old) are much stronger than younger ones. This strength increase with tectonic province age cannot be explained by conductive cooling of a lithospheric plate with a given thermal thickness as in oceanic lithosphere. For example, both the Sveconorwegian (~ 1 Gyr old) and Karelina (~ 3 Gyr old) tectonic provinces have had sufficient time (> 1 Gyr) to conductively cool and thermally equilibrate, yet T_e in the former is at least half that of the latter (Fig. 2). Therefore, there must be fundamental differences

in the mechanical structure between old and young continental lithosphere in Europe.

We suggest that these structural differences may reflect changes in continental plate forming processes related to the decrease in temperature and volatile content in the sublithospheric mantle during Earth's history. In the Archaean, higher mantle temperatures and/or volatile content probably favoured a larger degree of melting to greater depths than today and, hence, the formation of a thick lithosphere, with a highly depleted, buoyant root^{4,22}. At that time, the

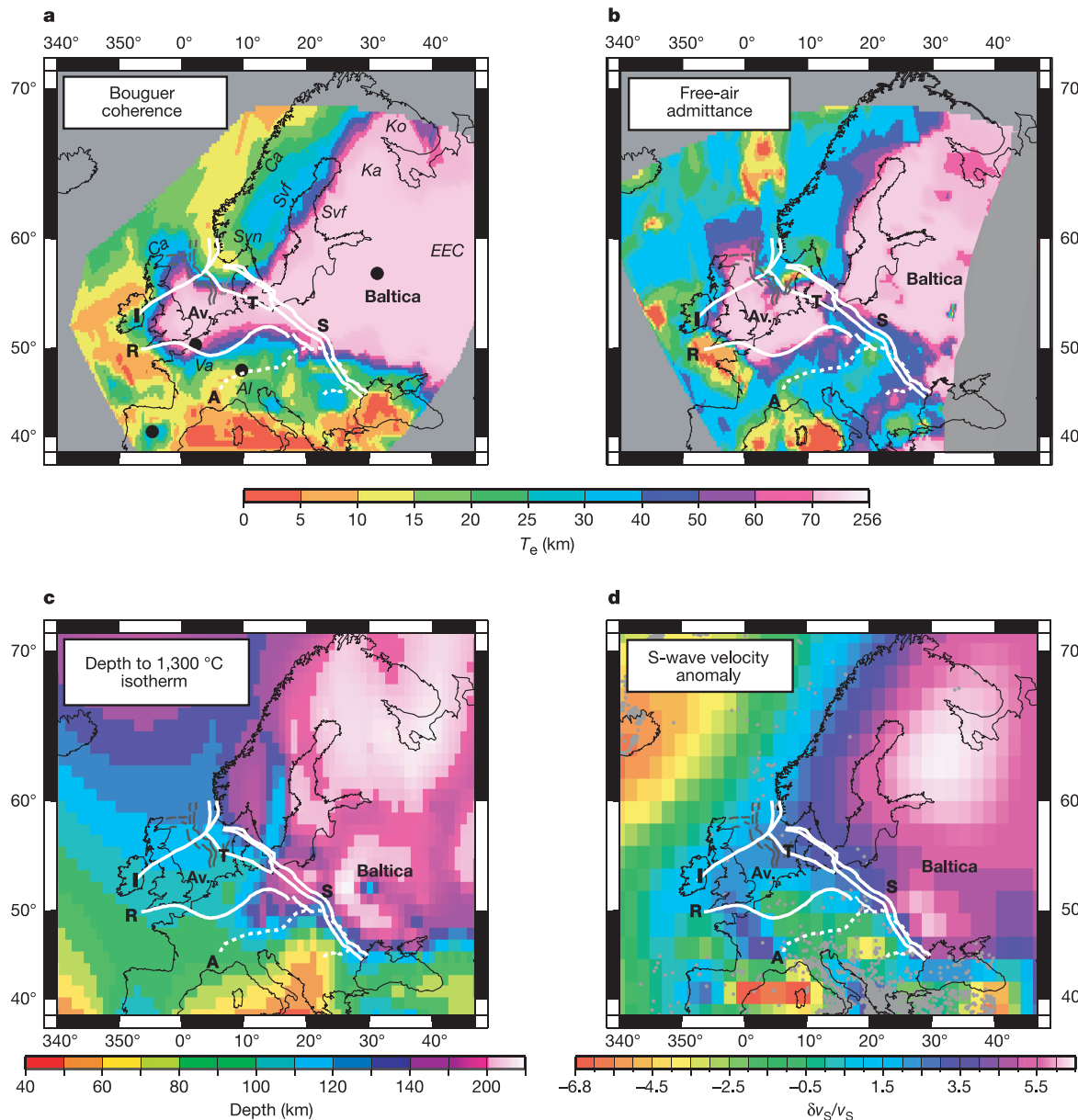


Figure 1 | T_e structure of Europe obtained using two different methods, and comparison with other geophysical data. a, b, T_e obtained from Bouguer coherence (a) and free-air admittance (b). c, d, Other geophysical data: c, thermal thickness³ (defined as depth to the 1,300 °C isotherm); and d, S-wave velocity anomaly, $\delta v_s/v_s$, at 100 km depth²⁰ given in per cent with respect to an isotropic version of PREM (described in ref. 20). Note that even if T_e values do not exactly coincide in a and b, the general pattern of T_e variation is equivalent. Free-air admittance has a poorer T_e recovery ability than the Bouguer coherence (Methods); we consider the results obtained with the latter more reliable. The T_e structure in a and b is based on CRUST 2.0³⁰, Poisson ratio = 0.25, and Young's modulus = 100 GPa. Because the maximum T_e that can be recovered with confidence is 60 km, larger T_e estimates exceed this value by an undetermined amount (Methods). Hence

the relative strength between the Precambrian Baltica and Avalonia (Av) can not be resolved. White lines define the sutures: I, Iapetus; T, Thor; R, Rheic; and S, Sorgenfrei-Tornquist and Teisseyre-Tornquist zones, between the Precambrian provinces and younger ones. Italic labels show the approximate location of the tectonic provinces: Ko, Kola; Ka, Karelia; Svf, Svecofennian; EEC, East European Platform; Svn, Sveconorwegian; Ca, Caledonian; Va, Variscan; and Al, Alpine. Dashed white line (A) shows the Alpine deformation front. Grey lines indicate the main boundary faults of the North Sea rift, which borders the high- T_e area in Avalonia. Grey dots in d are earthquake locations (US Geological Survey Earthquake Data Base, $M_b > 4.0$). Black dots in a show locations of Bouguer coherence analysis shown in Supplementary Information.

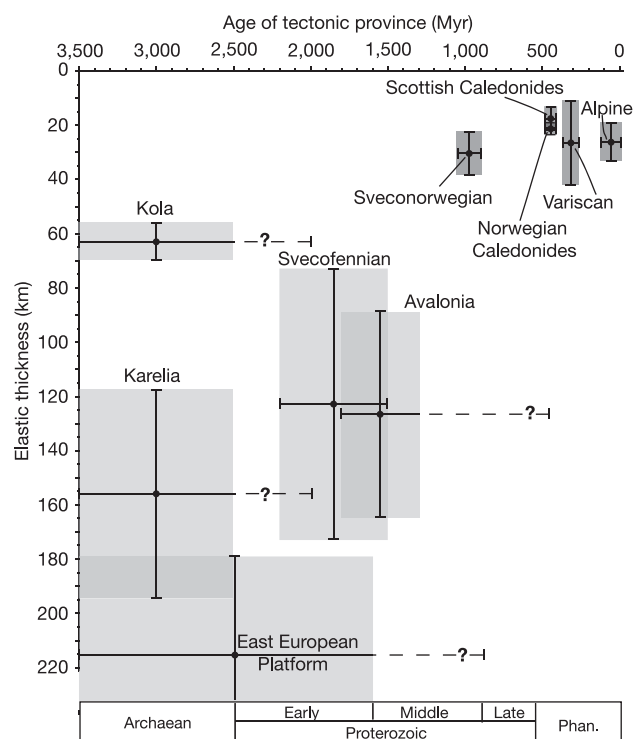


Figure 2 | Mean T_e determined from Bouguer coherence within each of the major tectonic provinces of Europe versus their age. Filled circles show data points; age is taken from Table 1. The age of the tectonic province corresponds to the age of those portions of the province that have not been deformed by later orogenic events at their edges. For example, the East European Platform consists mainly of Archaean and Early Proterozoic basement that are largely undeformed, so we assign this age instead of that of the later orogenies that have modified its edges. For tectonic provinces comprising only orogenic belts such as the Alps, we assign the age of the orogeny (see Table 1). To distinguish between tectonic province and orogenic age, we have coloured them in light and dark grey respectively. Error bars indicate the standard deviation in T_e and the age span of the tectonic provinces. Dashed error bars and question marks indicate uncertainty in the age of the tectonic province (see Table 1). As mentioned in Fig. 1, the maximum T_e estimate that we can recover is 60 km (Methods). Hence, variations in strength between tectonic provinces with $T_e > 60$ km cannot be interpreted.

response of the lithosphere to loading would have reflected a combination of weakening due to the high geothermal gradients and any strengthening due to compositional effects (related to melt extraction²²). These two competing factors might have resulted in an Archaean lithosphere, which despite its thickness, had a low strength and hence low T_e . However, with time, conductive cooling of the lithosphere to a stable state, as well as the Earth's secular cooling itself, would have increased its strength. This probably explains the high present-day T_e values that have been determined (both in this Letter and elsewhere) using spectral methods in old tectonic provinces (for example, North America^{6–8}, Australia⁹, South America²³ and Africa²⁴).

The decrease in temperature and volatile content in the sublithospheric mantle through time²² has resulted in Late Proterozoic and Phanerozoic lithosphere that is associated with smaller degrees of melting at shallower depths^{5,22}. Hence, this lithosphere is thinner⁵, has a higher geothermal gradient⁵, and is less depleted in basaltic constituents than Archaean lithosphere⁵. This makes Late Proterozoic and Phanerozoic lithosphere intrinsically weaker than older lithosphere, even after conductive cooling. This weakness is reflected, we believe, in the low T_e values obtained here in young tectonic provinces using spectral methods (Fig. 2).

It is interesting to note that the youngest ages for lithosphere

Table 1 | Age of European tectonic provinces

Tectonic province	Age range (Gyr)
Kola*	3.5–2.5
Karelia*	3.5–2.5
Svecofennian†	2.2–1.5
East European Platform‡	3.5–1.6
Avalonia§	1.3–1.8
Sveconorwegian	1.05–0.9
Caledonian	0.488–0.416
Variscan	0.359–0.27
Alpine	0.12–0

* The Kola and Karelian tectonic provinces comprise Archaean basement with a U–Pb age in the range 2.5–3.5 Gyr (ref. 13) and were amalgamated 2 Gyr ago. Hence, we assign them an age span of 3.5–2.5 Gyr. In Fig. 2 we extend the age error bar to 2 Gyr, as the extent to which the amalgamation 2 Gyr ago modified the Kola and Karelia lithospheres is uncertain.

† The Svecofennian tectonic province contains material younger than 2.2 Gyr and was accreted and underwent collision 2.0–1.8 Gyr ago. It was later locally reworked by melting 1.8–1.5 Gyr ago¹³. We assign an age range of 2.2–1.5 Gyr to this tectonic province.

‡ The East European Platform consists mainly of basement of Archaean and Early Proterozoic age and locally middle Proterozoic age³¹. We assign an age range of 3.5–1.6 Gyr. In Fig. 2 we extend the age error bar to 0.9 Gyr, with a question mark to indicate that the extent to which the local middle Proterozoic basement is represented in the T_e of this province is uncertain.

§ The Avalonia basement is structurally very heterogeneous and rarely exposed. Therefore, we have based our age range on Sm–Nd model ages, which vary from 1.3 to 1.8 Gyr (ref. 32). Note that in Fig. 2 we have extended the age error bar up to 443 Myr ago (when Avalonia docked against Baltica³³), with a question mark to indicate the uncertainty about the extent to which later tectonic events affecting Avalonia's edges modified its lithospheric core.

with seismic and thermal thicknesses greater than 200 km are around 1.6–1.7 Gyr (refs 2, 3), a similar age to that at which T_e in this study, appears to change markedly, except, perhaps, for Avalonia (Fig. 2). However, given that Avalonia is structurally very heterogeneous, its highest T_e values may reflect the areas where the basement ages are oldest (see Table 1).

Although it is not yet known exactly when cratons acquired their strength and stability^{12,25}, it is likely that, at least, they were strong throughout the Phanerozoic. If so, continental deformation would preferentially occur at the edges of stable cratonic provinces. Figure 1 shows some weakening of Baltica, for example, where it is juxtaposed to the Caledonian, Variscan and Alpine orogenic belts. We follow Dixon *et al.*²⁶ and speculate that water introduced during subduction may have further weakened the Phanerozoic tectonic provinces. The net result is a tectonic province that because of its weakness acts as a focus, such that sites of orogeny become repeatedly involved in both compressional and extensional deformation, as is predicted in the Wilson cycle.

Finally, our results indicate that the T_e of old tectonic provinces (≥ 1.5 Gyr) is significantly larger (> 60 km) than their mean crustal thickness (~ 40 km)²⁷. This suggests that the lithospheric mantle is strong, consistent with dynamical models that indicate that the stability of cratons is due not only to the chemical buoyancy of their root but also, importantly, to root strength^{28,29}. It appears that only when the cratonic lithosphere is subjected to processes such as enrichment by hot upwelling mantle will the old cratons be weakened enough to deform⁵.

METHODS

Calculations and data. The Bouguer coherence and free-air admittance are statistical methods that determine the relationship between the Bouguer and free-air gravity anomaly and the topography as a function of wavelength⁷. Both methods consist in finding a 'best fit' T_e by minimizing the root-mean-square difference between observed and predicted coherence and admittance functions⁷.

We base our analysis on a continent-wide 8×8 km grid of gravity anomaly (Bouguer onshore and free-air offshore) and topography data. These were compiled by GETECH (UK) as part of their West-East Europe Gravity Project (WEEGP). The gravity anomalies have been corrected for terrain, and we estimate that the data are accurate to better than 1–2 mGal. The calculation of the Bouguer anomaly offshore and the free-air anomaly onshore is described elsewhere¹¹.

The calculation of the Bouguer coherence and free-air admittance follows refs 7 and 11. The only differences from the methods followed in ref. 11 is that we:

(1) deconvolve the loads in the same data window as the observed functions, (2) assume that subsurface loads occur at the boundary of the upper and middle crust, and (3) combine the T_e results obtained using windows of different sizes. The model input parameters are summarized in Fig. 1.

The load deconvolution consists of extracting the contribution to the observed topography and gravity anomaly of the surface and subsurface loads that were initially emplaced on the lithosphere^{7,11}. The deconvolution requires information on the density structure of the crust, which we deduced from CRUST 2.0³⁰.

The subsurface loads occur at the upper/middle crust boundary, which, according to CRUST 2.0³⁰, is at ~10–15 km depth. Because the predicted gravity anomaly due to subsurface loading increases with T_e and decreasing depth of loading, any given gravity anomaly can be modelled by a combination of either a small loading depth and small T_e or a larger depth and larger T_e . However, we have found that the T_e variation resulting from different assumed loading depths is not large (± 5 km).

To recover a spatially varying T_e , the window size used needs to be large enough to recover the maximum flexural wavelength, but small enough to recover the spatial variation in T_e (Supplementary Fig. 3 illustrates how T_e varies with window size). To obtain an optimal solution, we used overlapping windows spaced 56 km apart and assigned the resulting T_e to the window centre. The study region was analysed four times using window sizes of 400×400 km, 600×600 km, 800×800 km and $1,000 \times 1,000$ km. We therefore generated our 'final' T_e structure using a combination of different window sizes. In particular, for $T_e < 20$ km, $20 < T_e < 40$ km, $40 < T_e < 60$ km and $T_e > 60$ km, we used the results obtained with windows of 400×400 km, 600×600 km, 800×800 km and $1,000 \times 1,000$ km, respectively.

Limitations. On the basis of tests with synthetic topography and gravity anomaly data, it has been found that, for a window of $1,000 \times 1,000$ km, the largest T_e that can be recovered with confidence is 60 km (ref. 11; see also Supplementary Information). Hence, the amount by which any particular T_e estimate exceeds this value is undetermined. In addition, in small areas where T_e is high (for example, central Finland), the use of large windows results in a reduction of T_e due to the inclusion in the analysis of low- T_e areas that flank the high- T_e areas. We therefore only place a lower limit on the largest T_e obtained (60 km), and attach little significance to local spatial variations of T_e within the Baltica and Avalonia tectonic provinces.

The performance of the free-air admittance in recovering T_e values is poorer than that of the Bouguer coherence owing to the relatively low power of the free-air anomaly at wavelengths where the isostatic compensation occurs (see, for example, ref. 11). Hence, we consider the results obtained with Bouguer coherence to be more reliable.

Received 21 January; accepted 18 May 2005.

- McConnell, R. B. Geological development of the rift system of Eastern Africa. *Geol. Soc. Am. Bull.* **83**, 2549–2572 (1972).
- Polet, J. & Anderson, D. L. Depth extent of cratons as inferred from tomographic studies. *Geology* **23**, 205–208 (1995).
- Artemieva, I. M. & Mooney, W. D. Thermal evolution of Precambrian lithosphere: A global study. *J. Geophys. Res.* **106**, 16387–16414 (2001).
- Jordan, T. H. Composition and development of the continental lithosphere. *Nature* **274**, 544–548 (1978).
- O'Reilly, S. Y., Griffin, W. L., Poudjom Djomani, Y. H. & Morgan, P. Are lithospheres forever? Tracking changes in subcontinental mantle through time. *GSA Today* **11**(4–10) (2001).
- Bechtel, T. D., Forsyth, D. W., Sharpton, V. L. & Grieve, R. A. F. Variations in the effective elastic thickness of the North American lithosphere. *Nature* **343**, 636–638 (1989).
- Forsyth, D. Subsurface loading and estimates of the flexural rigidity of continental lithosphere. *J. Geophys. Res.* **90**, 12623–12632 (1985).
- Audet, P. & Mareschal, J.-C. Variations in elastic thickness in the Canadian Shield. *Earth Planet. Sci. Lett.* **226**, 17–31 (2004).
- Swain, C. J. & Kirby, J. F. The effect of 'noise' on estimates of the elastic thickness of the continental lithosphere by the coherence method. *Geophys. Res. Lett.* **30**, 1574–1578, doi: 10.1029/2003GL017070 (2003).
- McKenzie, D. Estimating T_e in the presence of internal loads. *J. Geophys. Res.* **108**, doi:10.1029/2002JB001766 (2003).
- Pérez-Gussinyé, M., Lowry, A., Watts, A. B. & Velicogna, I. On the recovery of the effective elastic thickness using spectral methods: examples from synthetic data and the Fennoscandian shield. *J. Geophys. Res.* **109**, doi: 10.1029/2003JB002788 (2004).
- Grotzinger, J. & Royden, L. Elastic strength of the Slave craton at 1.9 Gyr and implications for the thermal evolution of the continents. *Nature* **347**, 64–66 (1990).
- Windley, B. in *A Continent Revealed. The European Geotraverse* (eds Blundell, D., Freeman, R. & Mueller, S.) 139–214 (Cambridge Univ. Press, Cambridge, UK, 1992).
- Kogan, M. G., Fairhead, J. D., Balmino, G. & Makedonskii, E. L. Tectonic fabric and lithospheric strength of northern Eurasia based on gravity data. *Geophys. Res. Lett.* **21**, 2653–2656 (1994).
- Poudjom Djomani, Y. H., Fairhead, J. D. & Griffin, W. L. The flexural rigidity of Fennoscandia: reflection of the tectonothermal age of the lithospheric mantle. *Earth Planet. Sci. Lett.* **174**, 139–154 (1999).
- Cloetingh, S. et al. Lithospheric memory, state of stress and rheology: neotectonic controls on Europe's intraplate continental topography. *Quat. Sci. Rev.* **24**, 241–305 (2005).
- Stewart, J. & Watts, A. B. Gravity anomalies and spatial variations of flexural rigidity at mountain ranges. *J. Geophys. Res.* **102**, 5327–5352 (1997).
- McBride, J. H., Snyder, D. B., England, R. W. & Hobbs, R. W. Dipping reflectors beneath old orogens: A perspective from the British Caledonides. *GSA Today* **6**, 1–6 (1996).
- Mona Lisa Working Group. Closure of the Tornquist Sea: constraints from MONA LISA deep reflection seismic data. *Geology* **22**, 617–620 (1997).
- Boschi, L., Ekstroem, G. & Kutowski, B. Multiple resolution surface wave tomography, the Mediterranean basin. *Geophys. J. Int.* **157**, 293–304 (2004).
- Watts, A. B. & Burov, E. Lithospheric strength and its relationship to the elastic and seismogenic layer thickness. *Earth Planet. Sci. Lett.* **213**, 113–131 (2003).
- Pollack, H. Cratonization and thermal evolution of the mantle. *Earth Planet. Sci. Lett.* **80**, 175–182 (1986).
- Ussami, N., Cogo de Sa, N. & Molina, E. C. Gravity map of Brazil. 2. Regional and residual anomalies and their correlation with major tectonic provinces. *J. Geophys. Res.* **98**, 2199–2208 (1993).
- Hartley, R., Watts, A. B. & Fairhead, J. D. Isostasy of Africa. *Earth Planet. Sci. Lett.* **137**, 1–18 (1996).
- Pearson, D. G. et al. Archaean Re-Os age for Siberian eclogites and constraints on Archaean tectonics. *Nature* **374**, 711–713 (1995).
- Dixon, J. E., Dixon, T. H., Bell, D. R. & Malservisi, R. Lateral variation in upper mantle viscosity: role of water. *Earth Planet. Sci. Lett.* **222**, 451–467 (2004).
- Christensen, N. I. & Mooney, W. D. Seismic velocity, structure and composition of the continental crust—global view. *J. Geophys. Res.* **100**, 9761–9788 (1995).
- Doin, M. P., Felitout, L. & Christensen, U. Mantle convection and the stability of depleted and undepleted continental lithosphere. *J. Geophys. Res.* **102**, 2771–2787 (1997).
- Lenardic, A., Moresi, L. N. & Mulhaus, L. N. Longevity and stability of cratonic lithosphere: Insight from numerical simulations of coupled mantle convection and continental tectonics. *J. Geophys. Res.* **108**, doi: 10.1029/2002JB001859 (2003).
- Laske, G., Masters, G. & Reif, C. The Reference Earth Model Website. (<http://mahi.ucsd.edu/Gabi/rem.html>) (2000).
- Goodwin, A. M. *Principles of Precambrian Geology* (Academic, San Diego, California, 1996).
- Brendan Murphy, J. & Nance, D. Sm-Nd isotopic systematics as tectonic tracers: an example from West Avalonia in the Canadian Appalachians. *Earth Sci. Rev.* **59**, 77–100 (2002).
- Wessel, P. & Smith, W. H. F. Free software helps map and display data. *Eos* **72**, 441–446 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Fairhead (GETECH, UK) for the provision of the gravity anomaly and topography data used in this Letter, J.-C. Mareschal and T. Lowry for constructive comments, and B. Holtzman, C. Mac Niocaill, S. Lamb, C. R. Ranero, J. Phipps Morgan, T. Jordan, T. Cunha, J. Hillier and G. Kozyreff for comments and discussions about the Letter. This work was supported by NERC. The figures presented here were constructed using GMT³³.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.P.-G. (martap@earth.ox.ac.uk and mperez@ija.csic.es).

Reinforcement of pre-zygotic isolation and karyotype evolution in *Agrodiaetus* butterflies

Vladimir A. Lukhtanov^{1*}, Nikolai P. Kandul^{2*}, Joshua B. Plotkin³, Alexander V. Dantchenko¹, David Haig² & Naomi E. Pierce²

The reinforcement model of evolution argues that natural selection enhances pre-zygotic isolation between divergent populations or species by selecting against unfit hybrids^{1,2} or costly interspecific matings³. Reinforcement is distinguished from other models that consider the formation of reproductive isolation to be a by-product of divergent evolution^{4,5}. Although theory has shown that reinforcement is a possible mechanism that can lead to speciation^{6–8}, empirical evidence has been sufficiently scarce to raise doubts about the importance of reinforcement in nature^{6,9,10}. *Agrodiaetus* butterflies (Lepidoptera: Lycaenidae) exhibit unusual variability in chromosome number. Whereas their genitalia and other morphological characteristics are largely uniform, different species vary considerably in male wing colour, and provide a model system to study the role of reinforcement in speciation. Using comparative phylogenetic methods, we show that the sympatric distribution of 15 relatively young sister taxa of *Agrodiaetus* strongly correlates with differences in male wing colour, and that this pattern is most likely the result of reinforcement. We find little evidence supporting sympatric speciation: rather, in *Agrodiaetus*, karyotypic changes accumulate gradually in allopatry, prompting reinforcement when karyotypically divergent races come into contact.

Speciation is the process whereby previously interbreeding populations develop reproductive isolation. Geographic barriers can arise and prevent gene flow, enabling populations to diverge genetically in allopatry^{4,11}. Occasionally, incipient allopatric species come into secondary contact through the expansion of their ranges before they have developed pre-zygotic isolating mechanisms. In such cases, natural selection acting against maladaptive hybrids^{1,2,6} or against costly interspecific mating^{6–8} can produce an indirect selection pressure favouring trait divergence and assortative mating. This process, called reinforcement, is of particular significance because it provides a role for natural selection in the formation of pre-zygotic isolation and eventually in speciation, processes that are otherwise incidental. Despite its plausibility⁷ and increasing attention from evolutionary biologists⁸, only a few well-documented cases of reinforcement have been published^{6,11–15}.

In the broad sense, reinforcement of pre-zygotic isolation can take place at both intraspecific and interspecific levels (see page 354 of ref. 3). Reinforcement between divergent populations that exchange genes⁹ can lead to speciation (termed “true reinforcement” by ref. 3), whereas reinforcement without gene flow is an adaptive genetic change that can occur after speciation has been completed. At both levels, reinforcement can give rise to a particular pattern of reproductive character displacement (RCD) involving greater interspecific mate discrimination between sympatric species than between allo-

patric species. Such patterns have long been considered evidence for reinforcement.

However, RCD is a common phenomenon⁵ and the same pattern of RCD may be generated by at least three other mechanisms: differential fusion^{9,16,17}, ecological character displacement^{6,15} and runaway sexual selection¹⁸. Under differential fusion, RCD arises as a by-product of evolution in allopatry. Those populations that have serendipitously evolved strong mating discrimination can persist in secondary sympatry, whereas those populations lacking such discrimination fuse and lose their distinctiveness. Species that persist in sympatry will demonstrate a high level of mating discrimination even though reinforcement has not operated^{11,16}. Similarly, ecological divergence may cause concomitant changes in mate recognition signals that make sympatric populations of two nascent species less likely to mate with one another^{6,15}. Runaway sexual selection can also generate RCD if selection has favoured dramatic differences in mate recognition characters directly within a single population arrayed along an ecological gradient¹⁸.

We have studied RCD in the species-rich genus *Agrodiaetus*. This genus is estimated to have arisen 2.5–3.8 million years ago¹⁹, and exhibits one of the widest diversities of chromosomal complements (that is, karyotypes) found in the animal kingdom, with haploid chromosome numbers of different species ranging from $n = 10$ to $n = 134$ (refs 20, 21). This range in karyotype is not caused by polyploidy: the similarity in genome sizes among *Agrodiaetus* species suggests that karyotype diversity arose through fusion and fragmentation of chromosomes^{19–21}. Hybrids between heterokaryotypic *Agrodiaetus* species have been observed in nature^{22,23}, but segregational problems during meiosis would result in their having reduced fertility. Karyotypic differences thus form a partial post-zygotic reproductive barrier^{20,24}. Although females are uniformly brown, *Agrodiaetus* species show considerable variability in male wing colour. Wing coloration, both in visible and ultraviolet wavelength ranges, is an important mate recognition characteristic in butterflies^{14,25} involved in the formation of pre-zygotic reproductive barriers^{14,15,26}. In lycaenids, both sexes typically exhibit mate choice²⁷, and females accept only those males with appropriate conspecific coloration²⁷. Females of polyommata species such as *Agrodiaetus* mate only once²⁸, and thus heterospecific matings that fail to give rise to viable offspring are strongly selected against. The combination of rapid karyotypic evolution, the role of karyotypic differences in reducing hybrid fitness, the reproductive biology of lycaenids, and a simple wing-colour-based criterion in mate choice makes *Agrodiaetus* a promising candidate for studies of reinforcement.

We reconstructed a phylogeny using 1,938 base pairs from two mitochondrial genes, *Cytochrome oxidase I* and *II* (*COI* and *COII*; see Supplementary Information), for 89 species and subspecies of

¹Department of Entomology, St Petersburg State University, Universitetskaya nab. 7/9, St Petersburg 199034, Russia. ²Department of Organismic and Evolutionary Biology, 26 Oxford Street, Harvard University, Cambridge, Massachusetts 02138, USA. ³Bauer Center for Genomics Research, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

*These authors contributed equally to this work.

Agrodiaetus. We determined the genetic distance on the phylogeny for each pair of sister taxa, and noted each taxon's wing colour and whether the taxa were sympatric or allopatric in their distribution. Traits involved in mate differentiation showed greater differences between sympatric pairs of species as opposed to allopatric pairs, when comparing taxa separated by relatively small genetic distances (ML distances ranging from 0–0.050 changes per nucleotide; see Supplementary Information).

By mapping taxon wing colour on the inferred phylogeny, we

observed 19 changes in wing colour among sampled taxa (Fig. 1). *A. cyaneus* and *A. gorbunovi* (Fig. 1) are the closest related pair of species found in sympatry (0.008 changes per nucleotide, under the HKY + I + Γ model of DNA substitution). Despite their genetic similarity, *A. cyaneus* has already acquired a new wing colour. Sampled sister taxa separated by genetic distances smaller than this value occur only in allopatry (Fig. 1). At genetic distances between 0.008 and 0.050, sympatric pairs of species begin to appear, and they generally exhibit different wing colours.

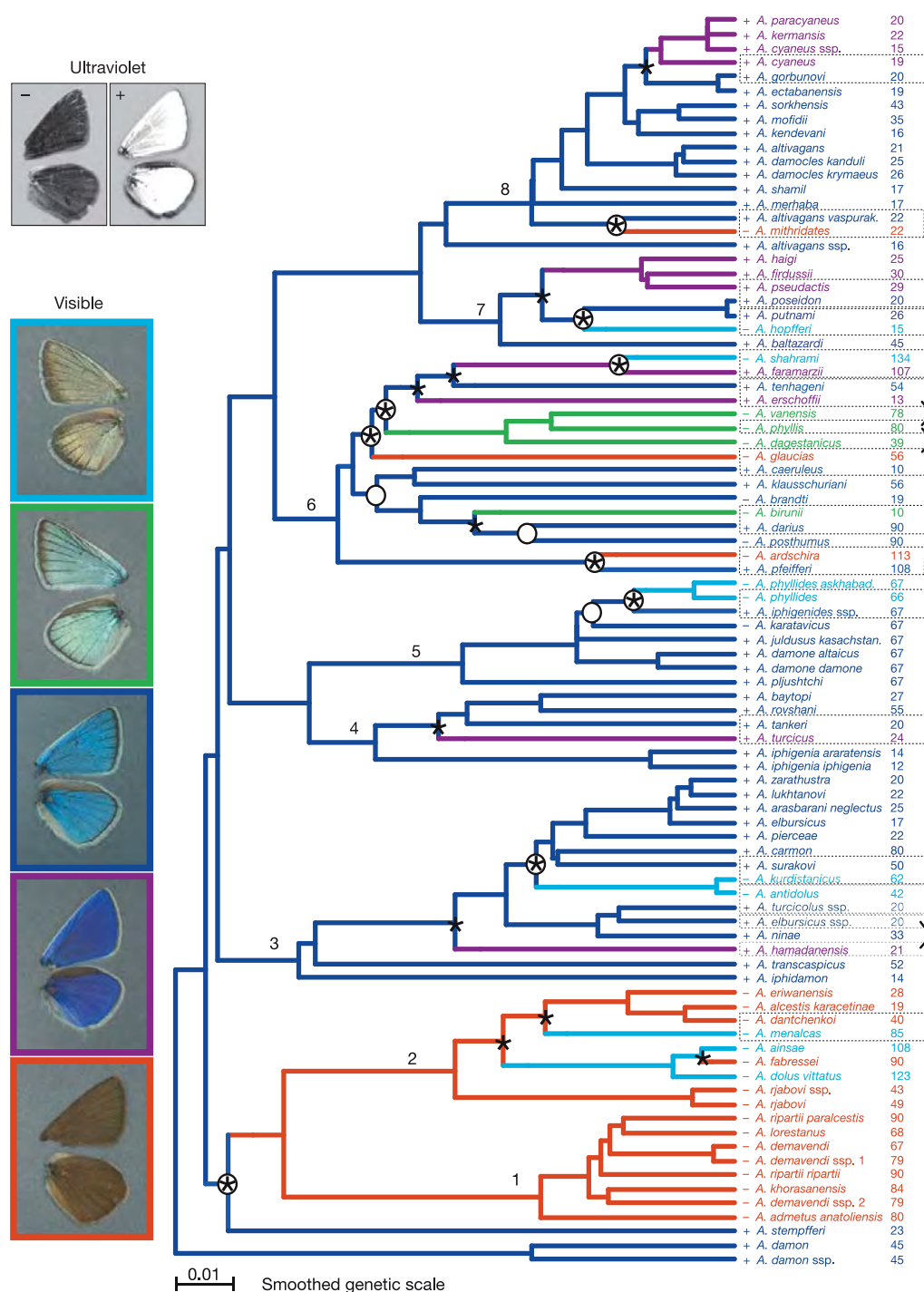


Figure 1 | Changes in male wing coloration along the phylogeny of *Agrodiaetus*. Five unordered states of male wing colour and presence (+) or absence (–) of wing ultraviolet reflectance were mapped on the ML tree. The inferred changes in wing visible coloration and ultraviolet reflectance are labelled on the tree with an asterisk and a circle, respectively. Relatively

young sympatric pairs of taxa with markedly different visible colours are boxed. The column to the right of the taxon names shows haploid chromosome numbers. Eight clades (1–8) were used to examine karyotypic divergence in *Agrodiaetus* (Fig. 4).

To account for dependencies arising from common ancestry, we analysed 88 independent pairs, or nodes, of sister clades²⁹. Of these nodes, 28 were classified as sympatric pairs, and the remaining 60 were allopatric. Out of 19 cases of colour changes (Fig. 1), 15 were observed between sympatric sister clades, and 4 between allopatric sister clades. In other words, colour changes occurred preferentially between sympatric, as opposed to allopatric, sister clades (Fisher's exact test, $P < 0.000002$). Considering only the sympatric pairs, most of the young sister clades (ML distances of 0–0.05 changes per nucleotide; Fig. 2a) exhibit colour differences, whereas old sister clades (0.05–0.12 changes per nucleotide; Fig. 2a) tend to have the same wing colour. Conversely, allopatric pairs of sister clades exhibit only a small number of colour differences, and these differences are distributed independently of genetic distance (Fig. 2b). A phylogenetic analysis of changes in wing ultraviolet reflectance produced a similar pattern of RCD (Figs 1 and 2; see also Supplementary Information).

Comparative phylogenetic methods applied to the geographic distributions of extant species enable us to discriminate between reinforcement and the three other mechanisms that could generate similar patterns of RCD. Reinforcement predicts that primarily young phylogenetic lineages will demonstrate RCD because older lineages are less likely to hybridize, having already acquired full reproductive isolation in allopatry. In contrast, differential fusion and ecological character displacement do not predict a strong correlation between RCD and lineage ages (see Supplementary Information). Moreover, according to differential fusion, RCD should be equally rare among old and young sympatric species, and changes in wing colour found among sympatric pairs of species should comprise a subset of the changes in colour seen among

allopatric pairs of species^{11,12}. The distribution of colour changes found in *Agrodiaetus* is therefore unlikely to be generated by differential fusion or ecological character displacement (Fig. 2), and is more consistent with the predictions of the reinforcement model.

The specific pattern of RCD found in this study (Fig. 2) could be caused by differential fusion if wing colour evolves rapidly, for example by sexual selection between allopatric populations, whereas other forms of pre-zygotic isolating characters are more stable in the genus *Agrodiaetus*. Our data argue against this possibility. Wing colour remained unchanged in 69 nodes among sampled *Agrodiaetus* taxa, and once evolved, new wing colours passed unchanged through multiple subsequent allopatric speciation events (Fig. 1). At the same time, other potentially pre-zygotic isolating characters such as host plant use and ecological preferences vary even between purely allopatric populations of *Agrodiaetus* (see Supplementary Information). The third alternative mechanism, runaway sexual selection, generates RCD within “a single population distributed across an ecological cline” (ref. 18; in sympatry), whereas primary divergence in allopatry is necessary for the appearance of RCD under reinforcement. In our data, the smallest genetic divergences occurred between sister taxa with allopatric distributions. An additional age–range correlation test³⁰ did not reveal a pattern consistent with frequent sympatric speciation in *Agrodiaetus* (Fig. 3). Thus, the relationships between genetic distance, male wing colour variability and geographic distribution exhibited by *Agrodiaetus* are consistent with reinforcement as a mechanism generating RCD, and appear to reject three alternative mechanisms: differential fusion, ecological character displacement and runaway sexual selection.

Eight independent clades (1–8; Fig. 1) were chosen to examine the accumulation of karyotypic diversity in the genus. These clades

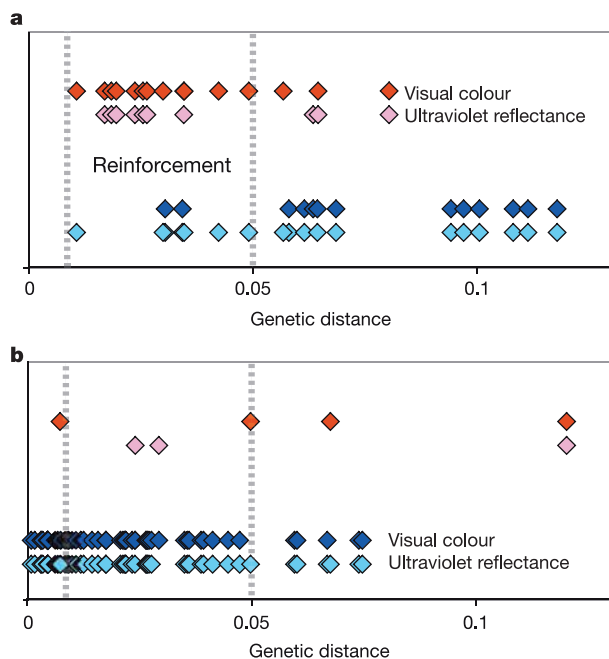


Figure 2 | Changes in male wing coloration between *Agrodiaetus* sister clades as a function of their genetic distance. According to our reconstruction, pairs of sister clades exhibit a change in visible colour at the node that separates them (red) or remain the same colour (blue). Ten changes in ultraviolet reflectance (pink) coincided with changes in visible coloration. Six changes in visible coloration between young sympatric sister clades did not affect wing ultraviolet reflectance (turquoise). **a**, Among sympatric sister clades, changes in visible colour occur primarily between recently divergent clades. **b**, Conversely, among allopatric sister clades, recently divergent clades retain the same coloration (visible and ultraviolet), and colour changes are otherwise rare and happen at random throughout the entire range of genetic distances.

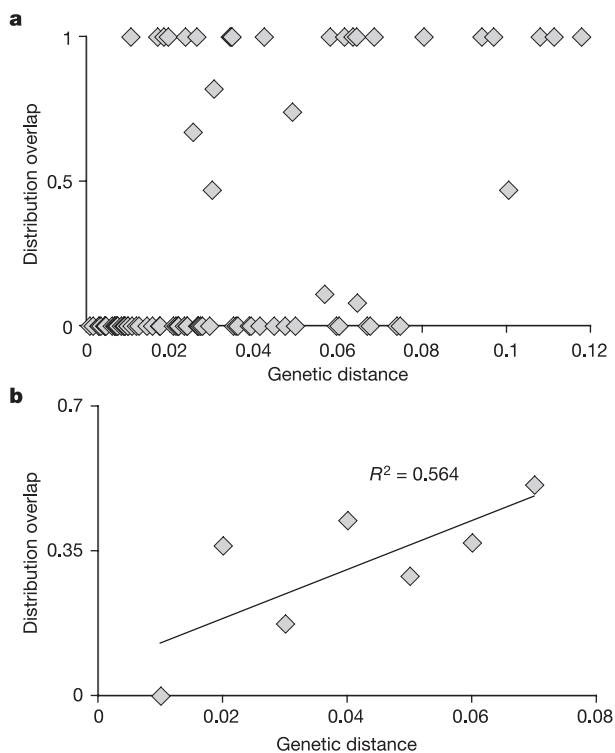


Figure 3 | Age-range correlation plot. **a**, The distribution overlap for every pair of sister clades was plotted against the genetic distance between them. **b**, The cumulative age-range correlation plot shows distribution overlap averaged over genetic distance. Because the number of pairs of relatively old sister clades was too small to calculate a mean distribution overlap, its values are not shown for genetic distances greater than 0.08 changes per nucleotide (under HKY + I + Γ).

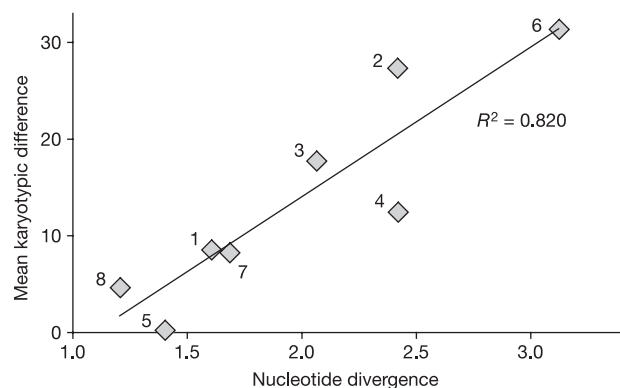


Figure 4 | *Agrodiaetus* karyotypic diversity strongly correlates with nucleotide divergence ($R^2 = 0.820$; $P < 0.002$). Mean intraclade karyotypic differences in eight independent clades (1–8, Fig. 2) are plotted against nucleotide divergences in the same clades.

exhibited a significant correlation between nucleotide divergence and mean karyotypic difference (Fig. 4; $R^2 = 0.820$; $P < 0.002$). Because genetic diversity among lineages is also proportional to the time since their divergence, it seems that chromosome numbers diverge with time. This conclusion is consistent with allopatric speciation as the main mode of cladogenesis in the genus, because karyotypic differences can accumulate within geographically isolated, small populations. Given the frequent chromosomal rearrangements in *Agrodiaetus*, karyotypic characters could act synergistically with geographic isolation to enhance reproductive barriers between nascent species of *Agrodiaetus*, despite their high vagility as adult butterflies. This synergistic action could impede gene flow between populations, facilitating the evolution of pre-zygotic isolating characters.

Although interspecific hybridization can occur within *Agrodiaetus*^{22,23}, we do not know whether nascent *Agrodiaetus* species exchange genes in sympatry before they acquire full reproductive isolation. Biological species can evolve pre-zygotic isolating barriers without gene flow, for example when their hybrids are sterile^{3,9}. At the same time, the absence of gene flow between extant sympatric species does not necessarily imply that these species did not exchange genes when they first came into contact. Therefore, although we cannot distinguish at what level (intraspecific or interspecific) reinforcement has operated, our comparative study demonstrates that natural selection against maladaptive matings is likely to have caused widespread divergence in pre-zygotic isolating characters between sympatric species of *Agrodiaetus*, and could have led to speciation.

METHODS

Methods are described in greater detail in the Supplementary Information.

Sampled species. *Agrodiaetus* butterflies belong to the section *Polymmatius* of the family Lycaenidae (Insecta: Lepidoptera). Females are brown, whereas males have a variety of background colours ranging from silver and blue to brown on the upper side of their wings (Fig. 1). With the exception of male wing coloration, which is a relatively labile character¹⁹ in the genus, *Agrodiaetus* species have remarkably similar genitalia and other external morphological characteristics. The taxa sampled for this study represent the entire range of known karyotypic diversity in *Agrodiaetus*, from $n = 10$ in *A. caeruleus* and *A. birunii* to $n = 134$ in *A. shahrami*. Identification of a number of *Agrodiaetus* species is based on karyotype; therefore the karyotypes of most specimens were examined before their DNA was extracted for gene sequencing, with the exception of 19 cases where individuals from populations with well characterized karyotypes were used^{20,21}. Eighty-nine well-differentiated taxa (76 species and 13 subspecies) of *Agrodiaetus* were used in this study (Supplementary Appendix 1), and of these, 52 karyotyped specimens of *Agrodiaetus* were analysed for the first time¹⁹.

Phylogenetic analysis. Two mitochondrial genes, *Cytochrome oxidase subunit I* (COI) and *Cytochrome oxidase subunit II* (COII), were amplified by polymerase chain reaction (PCR). PCR products were of equal length and directly

sequenced. Eighty-nine continuous sequences of COI, tRNA-leu and COII genes were aligned in a data set that was partitioned into the respective genes using PAUP* 4.0b10. For phylogeny reconstruction, we used three main methods: maximum parsimony (PAUP* 4.0b10), bayesian inference (MrBayes 3.01) and maximum likelihood (PHYML). Hierarchical likelihood ratio tests (hLRTs) were used to identify the model of DNA substitution that best fit the data for maximum likelihood and bayesian inference analyses. To ensure that the bayesian inference analysis was not trapped in local optima, we ran three independent rounds of the procedure. Average log-likelihood values at stationarity were calculated during each round and compared for convergence. The support of tree branches recovered by maximum parsimony and maximum likelihood methods was estimated with nonparametric bootstrap values. To align the tips of the recovered maximum likelihood tree (Fig. 1), we homogenized substitution rates across lineages using Sanderson's nonparametric rate-smoothing algorithm as implemented in TreeEdit.

Reconstruction of ancestral colour. Wing colour was treated as a multi-state unordered character with a total of five distinct states (Supplementary Appendix 1). Wing ultraviolet reflectance was coded as present or absent. A test of serial independence rejected the null hypothesis that the wing colour was not correlated with phylogeny ($P = 0.0003$). We used a maximum likelihood method of ancestral character reconstruction because this method accounts for branch lengths on the tree and estimates probabilities of reconstructing different states. We reconstructed ancestral wing coloration on the maximum likelihood tree (see Supplementary Information for branch support) inferred under the HKY + I + Γ model of DNA substitution (Fig. 1) in Mesquite 1.0. Maximum likelihood optimizations were done using the Markov k -state one-parameter model.

Sister-clade analysis. We compared reconstructed and extant states of species colours for every pair of sister clades on the maximum likelihood tree inferred under the HKY + I + Γ model of DNA substitution (Fig. 1; see Supplementary Information for branch support). Average genetic distances between sister clades were estimated from the maximum likelihood tree. We classified a pair of sister clades as sympatric if they shared at least one pair of basal taxa with a sympatric distribution (Supplementary Appendix 2). We have considered all nodes independent of their age. For simplicity, we have assumed that sister clades separated by relatively older nodes gained sympatry only recently. The observed pattern of RCD is even stronger if we exclude sister taxa separated by older nodes from our analysis.

Karyotype evolution. A test of serial independence ($P = 0.0003$) showed that the distribution of chromosome number is correlated with phylogenetic placement on the maximum likelihood tree. Although karyotype (including chromosome number and relative size of bivalents) is a labile character in *Agrodiaetus*^{19,21}, we must control for changes attributable to common ancestry. We chose eight independent clades (that is, lineages), recovered on the maximum likelihood tree (Fig. 1), to examine the accumulation of karyotypic divergence in *Agrodiaetus*. Nucleotide divergence in a clade was estimated in Arlequin. To calculate mean intraclade karyotypic differences, we first averaged haploid chromosome numbers between every pair of sister clades from the maximum likelihood tree, starting from the tips and working towards the root of the tree, took the absolute difference between averaged chromosome numbers at every node on the tree, and then averaged these differences at the internal nodes included in the eight well-defined clades (Fig. 1).

Received 5 February; accepted 4 May 2005.

1. Dobzhansky, T. Speciation as a stage in evolutionary divergence. *Am. Nat.* **74**, 312–321 (1940).
2. Noor, M. A. F. Speciation driven by natural selection in *Drosophila*. *Nature* **375**, 674–675 (1995).
3. Coyne, J. A. & Orr, A. H. *Speciation* (Sinauer Associates, Sunderland, Massachusetts, 2004).
4. Mayr, E. *Population, Species, and Evolution; an Abridgment of Animal Species and Evolution* (Harvard Univ. Press, Cambridge, Massachusetts, 1970).
5. Turelli, M., Barton, N. & Coyne, J. Theory and speciation. *Trends Ecol. Evol.* **16**, 330–343 (2001).
6. Noor, M. A. F. Reinforcement and other consequences of sympatry. *Heredity* **83**, 503–508 (1999).
7. Kirkpatrick, M. & Ravigne, V. Speciation by natural and sexual selection: models and experiments. *Am. Nat.* **159**, S22–S35 (2002).
8. Servedio, M. R. & Noor, M. A. F. The role of reinforcement in speciation: theory and data. *Annu. Rev. Ecol. Syst.* **34**, 339–364 (2003).
9. Butlin, R. K. Species, speciation, and reinforcement. *Am. Nat.* **130**, 461–464 (1987).
10. Marshall, J. L., Arnold, M. L. & Howard, D. J. Reinforcement: the road not taken. *Trends Ecol. Evol.* **17**, 558–563 (2002).

11. Coyne, J. A. & Orr, H. A. Patterns of speciation in *Drosophila*. *Evolution* **43**, 362–381 (1989).
12. Coyne, J. A. & Orr, A. H. "Patterns of speciation in *Drosophila*" revisited. *Evolution* **51**, 295–303 (1997).
13. Sætre, G.-P. *et al.* A sexually selected character displacement in flycatchers reinforces premating isolation. *Nature* **387**, 589–592 (1997).
14. Jiggins, C. D., Naisbit, R. E., Coe, R. L. & Mallet, J. Reproductive isolation caused by colour pattern mimicry. *Nature* **411**, 302–305 (2001).
15. Nosil, P., Crespi, B. J. & Sandoval, C. P. Reproductive isolation driven by combined effects of ecological adaptation and reinforcement. *Proc. R. Soc. Lond. B* **270**, 1911–1918 (2003).
16. Templeton, A. R. Mechanisms of speciation – a population genetic approach. *Annu. Rev. Ecol. Syst.* **12**, 23–48 (1981).
17. Butlin, R. K. Reinforcement: an idea evolving. *Trends Ecol. Evol.* **10**, 432–434 (1995).
18. Day, T. Sexual selection and the evolution of costly female preference: spatial effects. *Evolution* **54**, 715–730 (2000).
19. Kandul, N. P. *et al.* Phylogeny of *Agrodiaetus* Hübner 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA sequences of *COI* and *COII*, and nuclear sequences of *EF1- α* : karyotype diversification and species radiation. *Syst. Biol.* **53**, 278–298 (2004).
20. Lesse, de H. Spéciation et variation chromosomique chez les Lépidoptères Rhopalocères. *Ann. Sci. Nat. Zool. Biol. Anim.* **2**, 1–223 (1960).
21. Lukhtanov, V. A. & Danchenko, A. D. Principles of the highly ordered arrangement of metaphase I bivalents in spermatocytes of *Agrodiaetus* (Insecta, Lepidoptera). *Chromosome Res.* **10**, 5–20 (2002).
22. Hagen, W. T. Freilandhybriden bei Bläulingen aus Ostanatolien und Iran (Lepidoptera: Lycaenidae). *Nachr. entomol. Ver. Apollo* **23**, 199–203 (2003).
23. Schurian, K. G. & Hofmann, P. Ein neuer Lycaeniden-Hybrid: *Agrodiaetus ripartii* Freyer x *Agrodiaetus menalcas* Freyer (Lepidoptera: Lycaenidae). *Nachr. entomol. Ver. Apollo* **1**, 21–23 (1980).
24. Lorkovic, Z. in *Butterflies of Europe* (ed. Kudrna, O.) 332–396 (Aula, Wiesbaden, 1990).
25. Fordyce, J. A., Nice, C. C., Forister, M. L. & Shapiro, A. M. The significance of wing pattern diversity in the Lycaenidae: mate discrimination by two recently diverged species. *J. Evol. Biol.* **14**, 871–879 (2002).
26. Vane-Wright, R. I. & Boppre, M. Visual and chemical signalling in butterflies: functional and phylogenetic perspective. *Phil. Trans. R. Soc. Lond. B* **340**, 197–205 (1993).
27. Bernard, G. D. & Remington, C. Color vision in *Lycaena* butterflies: spectral tuning of receptor arrays in relation to behavioural ecology. *Proc. Natl Acad. Sci. USA* **88**, 2783–2787 (1991).
28. Drummond, B. A. in *Sperm Competition and the Evolution of Animal Mating Systems* (ed. Smith, R. L.) 291–370 (Academic, San Diego, California, 1984).
29. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985).
30. Barraclough, T. G. & Vogler, A. P. Detecting the geographic pattern of speciation from species-level phylogeny. *Am. Nat.* **155**, 419–434 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. J. Berry, J. A. Coyne, S. V. Edwards, J. R. Morris and R. Vila for their advice during the preparation of this manuscript. C. Bilgin, J. Coleman, F. Fernández-Rubio, G. Grigorjev, J. Jubany, C. Ibáñez, R. Martínez, M. L. Munguira, C. Sekercioglu, C. Stefanescu, M. A. Travassos, R. Vila and V. Zurilina helped with collecting specimens, and J. Coleman and R. Vila assisted with sequencing in the laboratory. W. H. Piel and A. Monteiro helped us to measure wing ultraviolet reflectance. This research was supported by three collecting grants from the Putnam Expeditionary Fund of the Museum of Comparative Zoology, Harvard University; a National Science Foundation Doctoral Dissertation Improvement Grant to N.P.K.; National Science Foundation and Baker Foundation grants to N.E.P.; Milton Fund grants to D.H. and J.B.P.; a Burroughs Wellcome Fund grant to J.B.P.; and grants from the Russian Foundation for Basic Research, and the Russian Federal Programs "Universities of Russia" and "Leading Scientific Schools" to V.A.L.

Author Information The sequences have been deposited in GenBank; see Supplementary Appendix 1 for details. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to N.E.P. (npierce@fas.harvard.edu).

LETTERS

Deep sub-seafloor prokaryotes stimulated at interfaces over geological time

R. John Parkes¹, Gordon Webster^{1,2}, Barry A. Cragg¹, Andrew J. Weightman², Carole J. Newberry², Timothy G. Ferdelman³, Jens Kallmeyer^{3†}, Bo B. Jørgensen³, Ivano W. Aiello⁴ & John C. Fry²

The sub-seafloor biosphere is the largest prokaryotic habitat on Earth¹ but also a habitat with the lowest metabolic rates². Modelled activity rates are very low, indicating that most prokaryotes may be inactive or have extraordinarily slow metabolism². Here we present results from two Pacific Ocean sites, margin and open ocean, both of which have deep, subsurface stimulation of prokaryotic processes associated with geochemical and/or sedimentary interfaces. At 90 m depth in the margin site, stimulation was such that prokaryote numbers were higher (about 13-fold) and activity rates higher than or similar to near-surface values. Analysis of high-molecular-mass DNA confirmed the presence of viable prokaryotes and showed changes in biodiversity with depth that were coupled to geochemistry, including a marked community change at the 90-m interface. At the open ocean site, increases in numbers of prokaryotes at depth were more restricted but also corresponded to increased activity; however, this time they were associated with repeating layers of diatom-rich sediments (about 9 Myr old). These results show that deep

sedimentary prokaryotes can have high activity, have changing diversity associated with interfaces and are active over geological timescales.

Recently, subsurface prokaryotes have been found to be ubiquitous on Earth (for example, in sediments, rocks, aquifers, mines, basalts and crustal fluids, oil reservoirs and ice sheets³). In the dark and remote from photosynthetically produced organic matter these environments are among the lowest-energy-flux habitats known⁴, with metabolic rates 10^3 – 10^5 times lower than in near-surface sediments⁵. Their enormous prokaryotic biomass¹ has therefore been questioned, particularly for the largest habitat, namely sub-seafloor sediments². In certain terrestrial habitats higher prokaryotic activity occurs at deep interfaces such as sandstone-shale⁶ or sandstone-clay⁷, showing active prokaryotic metabolism over geological timescales (for example the Cretaceous period⁶). Although deep stimulation of prokaryotic activity occurs in some sub-seafloor sediments^{8,9}, the impact of interfaces has not been fully explored, including biodiversity changes expected in dynamic prokaryotic

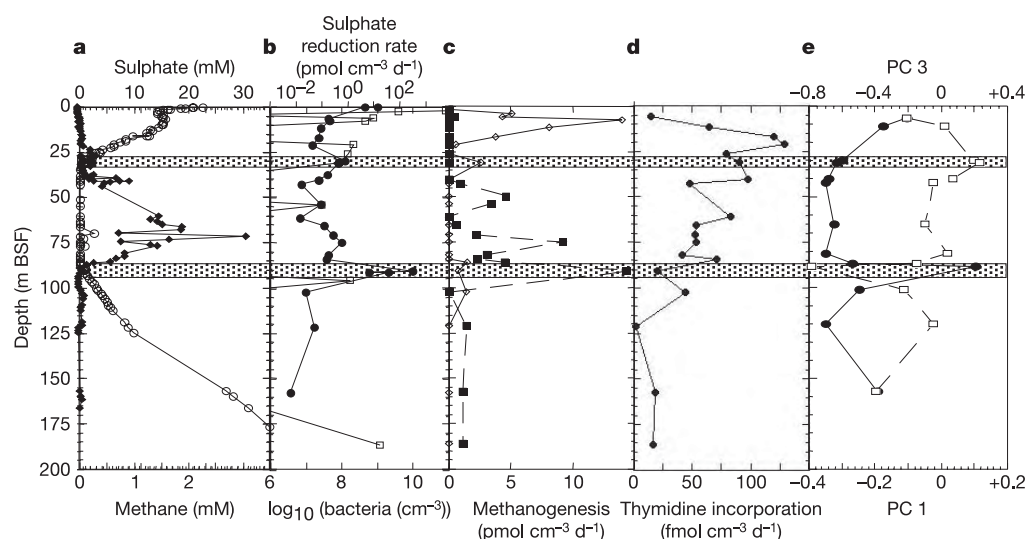


Figure 1 | Biogeochemical process and prokaryotic biodiversity profiles at the Peru margin site (ODP 1229). **a**, Geochemistry²⁹: open circles, pore water sulphate; filled diamonds, methane. **b**, Open squares, sulphate reduction rates; filled circles, total population of prokaryotes. **c**, Methanogenic rates: Open diamonds, H_2/CO_2 ; filled squares, acetate. **d**, Growth rates, measured as thymidine incorporation. **e**, Principal-

components profile of diversity of Bacteria from a DGGE analysis of 16S rRNA gene sequences: filled circles, component 1 (56% of variation); open squares, component 3 (9% of variation). Component 2 (24% of variation) had a similar profile to that of component 1. Shaded boxes highlight elevated prokaryotic processes and sulphate/methane interfaces.

¹School of Earth, Ocean and Planetary Sciences, Cardiff University, Main Building, Park Place, Cardiff CF10 3YE, UK. ²Cardiff School of Biosciences, Cardiff University, Main Building, Park Place, P.O. Box 915, Cardiff CF10 3TL, UK. ³Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany. ⁴Moss Landing Marine Laboratories, 8272 Moss Landing Road, Moss Landing, California 95039-9647, USA. [†]Present address: Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island 02882, USA.

communities adapted to local environments¹⁰. Prokaryotic stimulation and adaptation has been shown for shallow (4 m) marine layers that are high in organic matter¹¹ and suggested in deeper (58 m) clays with volcanic ash layers in the absence of geochemical or activity measurements¹². Here we investigated changes in prokaryotic activity, population size and composition in deep marine sediments (more than 400 m) from the east Pacific Ocean with deep geochemical or lithological interfaces¹³.

Sediments were from a continental margin and open ocean site (water depths of 150.5 and 3,297 m, respectively; Ocean Drilling Program (ODP) Leg 201). The margin site was unusual in having a deep brine incursion, so sulphate was present both in the near surface, from sea water, and at depth (Fig. 1a). A high content of organic matter in sediment (2–8% (ref. 14)) greatly stimulated prokaryotic activity, resulting in high populations of prokaryotes ($6.5 \times 10^8 \text{ cm}^{-3}$) and high rates of anaerobic reduction of sulphate (about $6,000 \text{ pmol cm}^{-3} \text{ d}^{-1}$; Fig. 1b) with sulphate reaching zero at about 35 m below sea floor (BSF). Below this depth, biogenic methane increased to about $2,000 \text{ } \mu\text{M}$ between 65 and 75 m BSF.

Separation between zones of sulphate and methane is normally interpreted as sulphate-reducing bacteria outcompeting methanogens for limiting substrates⁵, and hence the restriction of methanogenesis to deeper, sulphate-free sediments. However, our results show methanogenesis (about $15 \text{ pmol cm}^{-3} \text{ d}^{-1}$; Fig. 1c) within the sulphate zone, so in high-organic-matter sediments active methanogenesis can coexist with sulphate reduction¹⁵. The absence of methane in this zone (Fig. 1a) is therefore probably due to methane consumption, presumably by an anaerobic consortium of methane oxidizers and sulphate-reducing bacteria¹⁶. This process is normally intensified at the sulphate/methane interface¹⁷ and probably accounts for the increase in prokaryotic populations ($P < 0.001$), numbers of dividing and divided cells (data not shown) at about 30 m BSF (6.3-fold and 1.3-fold increase, respectively) and sulphate reduction rates (above and at the interface). There is also an increase in rates of H_2/CO_2 methanogenesis. Prokaryotic stimulation is repeated and intensified at the lower sulphate/methane interface, about 90 m, where total populations and rates of methanogenesis increase markedly (61-fold and 31-fold, respectively; $P < 0.001$). In addition there is a peak of sulphate reduction at the bottom of the interface (about 95 m BSF). At this interface prokaryotic populations are considerably larger (13-fold; $P < 0.001$) and rates of methanogenesis are comparable to those near the surface. Further, acetate becomes the major methanogenic substrate, in contrast to H_2/CO_2 methanogenesis near the surface; there is therefore a shift in carbon flow. Increases in acetate methanogenesis occur at depth at other sites and was related to warming and activation of organic matter with depth⁸. However, increased methane oxidation also occurs¹⁸ and this might result in increased acetate formation, either as a direct intermediate of anaerobic methane oxidation or as a product of recycling dead cells associated with greater prokaryotic biomass.

Prokaryotic growth, measured by thymidine incorporation, is greater in the subsurface than at the surface, with a maximum above the top sulphate/methane interface (Fig. 1d). In the methane zone, thymidine incorporation gradually decreases and reaches low levels that are maintained in the deep sulphate zone. There is no correlation between thymidine incorporation and other prokaryotic measurements, which might reflect the inability of some prokaryotes, including many sulphate-reducing bacteria and methanogens, to incorporate thymidine¹⁹. Thymidine incorporation is therefore probably reflecting an actively growing subset of the total population.

Gene libraries of 16S rRNA were obtained from high-molecular-mass DNA from all four depths (6.7, 30.2, 42.03 and 86.67 m BSF) although, as expected, extractable DNA yields were low^{20,21}. However, diversity in Bacteria was high, with sequences from at least seven major phyla or class-level groups (Fig. 2a, see also Supplementary Information). Gammaproteobacteria were numerous in the

shallowest (6.7 m BSF) and dominated the deepest (near the 90-m interface) libraries, whereas green non-sulphur (GNS) bacteria, although abundant throughout, dominated the libraries within the methane zone (30.2 and 42.03 m BSF). There was therefore a distinctly different bacterial community in the near-surface and

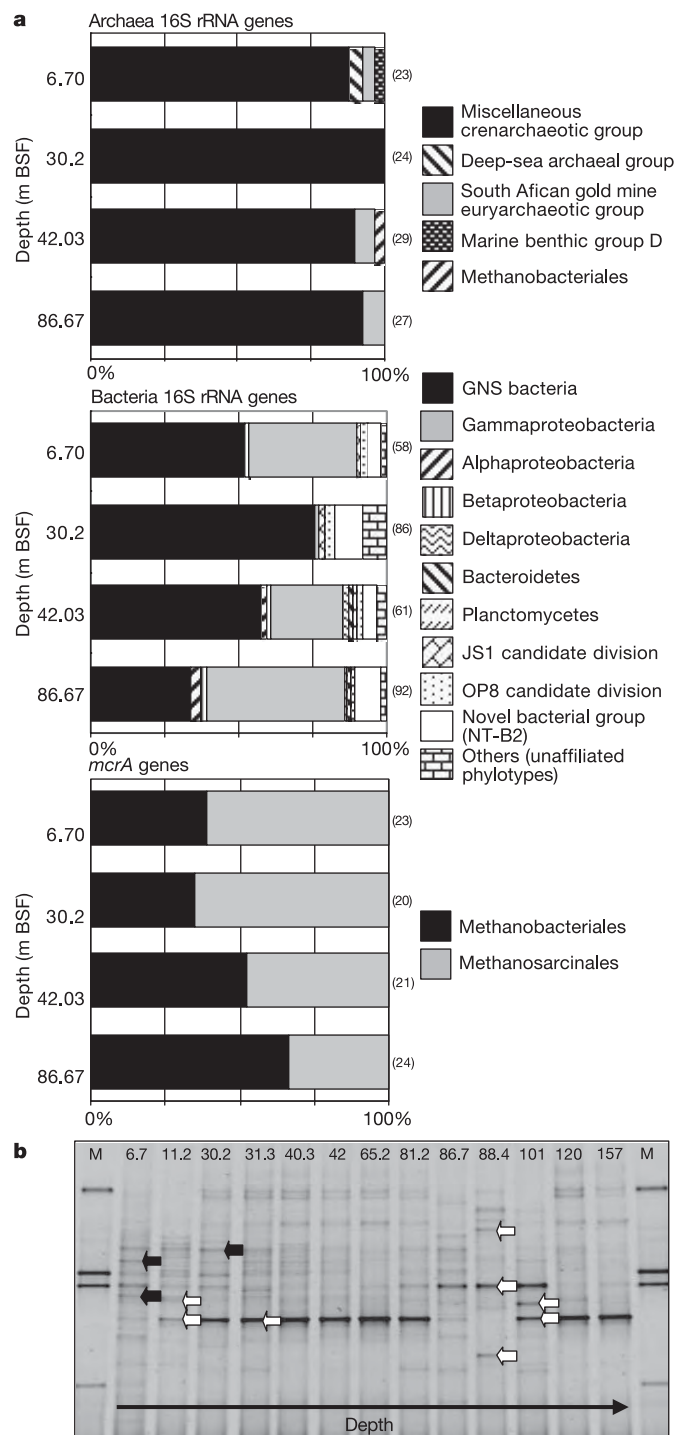


Figure 2 | Prokaryotic biodiversity at the Peru margin site (ODP Leg 201, site 1229). **a**, Biodiversity based on gene libraries for Archaea, Bacteria (16S rRNA gene) and methanogens (*mcrA* gene) from four depths. Numbers in parentheses are the numbers of sequenced clones in each library. **b**, DGGE profile of Bacteria from 13 depths (depth in metres BSF labelled at the top of individual lanes; M, marker lane²⁰). The identities of sequences from representative bands are indicated by arrows: black arrows, GNS bacteria; white arrows, Gammaproteobacteria.

deeper sulphate zones from that in the methane zone, and this corresponds to the zones of different thymidine incorporation (Fig. 1d). A new phylum widely distributed in anoxic sediments, JS1 (ref. 22), was only in the 6.7 and 30.2 m BSF libraries and was a minor component. This is unlike Nankai Trough sediments, in which JS1 dominated and no GNS bacteria were found²¹. However, it is similar to the Sea of Okhotsk deep sediments where Gammaproteobacteria dominated some layers, with GNS and JS1 Bacteria dominating others¹².

Denaturing gradient-gel electrophoresis (DGGE) was used to investigate 16S rRNA gene diversity at nine additional depths; this confirmed high diversity, with 24 distinct bands overall (Fig. 2b). Sequencing representative bands confirmed that GNS and Gammaproteobacteria were most abundant, although the new JS1 phylum was detected at all depths with targeted primers²². Principal-component (PC) analysis of DGGE depth profiles (three components comprised 79% of variation) showed distinct changes in biodiversity (Fig. 1e). PC1 and PC2 (70% of variation) had similar depth profiles, with abrupt changes very close to the 90 m interface (88.4 m BSF), whereas PC3 (9% of variation) had a contrasting profile, with positive values in the top 40 m BSF, peaking near the 30 m interface, and then decreasing. However, near the 90 m interface values changed rapidly to a minimum negative value and then increased below. Hence, there are abrupt changes in diversity of Bacteria at the two sulphate/methane interfaces, with populations at each being distinctly different (Fig. 2). The population at the 90 m interface is unique in containing only three bright bands, which were not present together at any other depth. PC1 is strongly positively correlated ($P < 0.05$) with dissolved sulphate and rates of H_2/CO_2 methanogenesis, whereas PC3 is positively correlated with alkalinity and thymidine incorporation but negatively with dissolved sulphate and manganese. This provides a unique insight into the potential characteristics and environmental control of prokaryotic diversity at this site.

We have demonstrated the stimulation of prokaryotic processes at discrete interfaces in deep, sub-seafloor sediments and shown marked changes in biodiversity and biogeochemical processes at subsurface interfaces. These results provide compelling evidence for active and dynamic populations in subsurface marine sediments

and are consistent with recent evidence for the viability of microscopically detected cells²³. Sediments at about 90 m BSF date to the early Pleistocene epoch (about 0.8 Myr ago), whereas the deeper sediments (186 m BSF) are Late Pliocene in age (up to 2 Myr ago²⁴). Prokaryotic processes are therefore operating on geological time-scales. In addition, this site was studied 18 years ago on ODP Leg 112 (ref. 25) and provided evidence, controversial at the time, for the stimulation of culturable prokaryotes below about 50 m BSF; however, these data now also show the long-term stability of stimulated subsurface prokaryotic populations.

In contrast to Bacteria, the diversity of Archaea was more limited, and as in other subsurface sediments^{12,21} all depths were dominated by the diverse miscellaneous crenarchaeotic group (Fig. 2a, see also Supplementary Information). Only one methanogen sequence was detected and this was from the methane zone (42.03 m BSF). However, consistent with the methanogenic rate measurements was the observation that methanogen-specific genes (*mcrA*) were present at all four depths (Fig. 2a). Diversity was limited to Methanobacteriales and Methanosarcinales (taxa using H_2/CO_2 and/or acetate, respectively), which is consistent with methane formation from both these substrates (Fig. 1c). Similar limited methanogenic diversity occurs in other deep sediments^{21,26}. Sequences for anaerobic methane oxidizers¹⁶ were absent. However, neither were sulphate-reducing bacteria detected, despite high (near-surface) rates of sulphate reduction, detectable sulphate reduction at depth (Fig. 1) and presumably sulphate reduction coupled to methane oxidation at the two sulphate/methane interfaces¹⁷. Calculations from the sulphate reduction rates indicate that the maximum proportion of sulphate-reducing bacteria of the total population at the 30 and 90 m interfaces is 0.02% and 0.002%, respectively, and it is most unlikely that these would be in our 16S rRNA gene libraries²⁷. Furthermore, low numbers of sulphate-reducing bacteria are consistent with a lack of DNA amplification with a specific gene from sulphate-reducing bacteria (*dsrAB* (ref. 28)). Low numbers could also occur with anaerobic methane-oxidizing prokaryotes, especially when associated with sulphate-reducing bacteria¹⁶. Hence, prokaryotes directly involved in sulphate reduction and/or anaerobic methane oxidation at this site may be highly active but might represent a small proportion of the total population or be unknown prokaryotes in

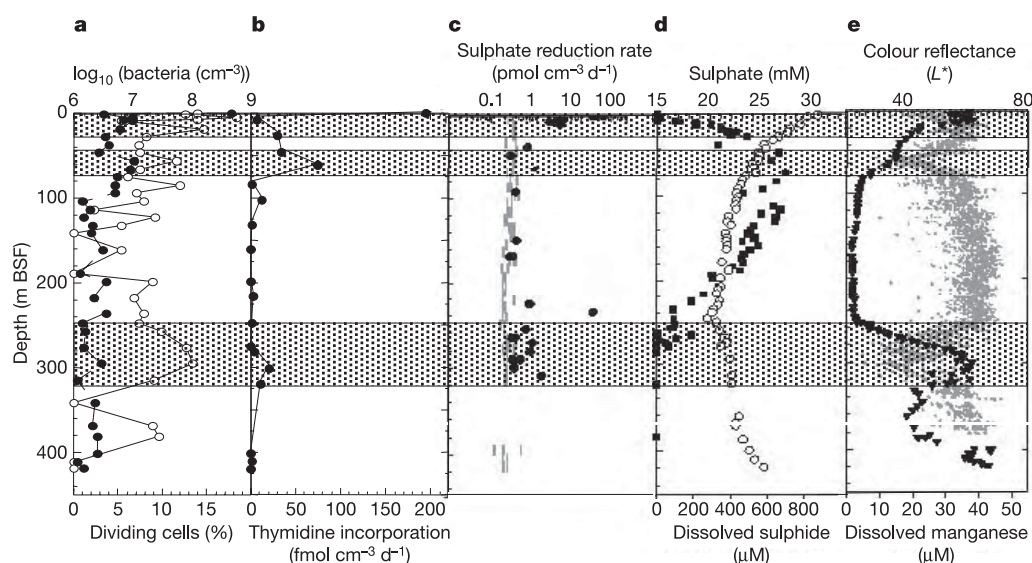


Figure 3 | Biogeochemical processes and prokaryotic populations at the Pacific open ocean site (ODP 1226). **a**, Prokaryotic profiles: filled circles, total prokaryotic population; open circles, percentage of dividing and divided cells. **b**, Rate of prokaryotic growth, measured as thymidine incorporation. **c**, Filled circles, sulphate reduction rates; grey oblongs,

minimum detection limits. **d**, Geochemistry²⁹: open circles, pore water sulphate; filled squares, hydrogen sulphide. **e**, Colour reflectance as a measure of diatom abundance (low reflectance represents high diatom abundance); filled triangles, pore water manganese concentration. Shaded boxes highlight elevated prokaryotic processes and high diatom layers.

our gene sequences, or their genes might not have been amplified.

Similar stimulation of prokaryotes occurred in the open ocean site but in association with repeated lithological depth changes and their allied high diatom content (Fig. 3). In the three diatom-rich layers between the surface and about 400 m BSF there was a consistent stimulation of prokaryotic activity (sulphate reduction and thymidine incorporation) and total prokaryotic numbers and/or the proportion of dividing and divided cells. Furthermore, in the top and bottom diatom-rich layers there was an increasing concentration of dissolved manganese²⁹, indicating active manganese reduction by prokaryotes. The mechanism of prokaryotic stimulation in these layers is not clear but it might be that diatomaceous organic matter is considerably less reactive than other sedimentary organic matter and as a consequence can fuel low, but continuing, prokaryotic activity over long periods. The deepest layer (about 250–320 m BSF) is 7–11 Myr old (ref. 29), which markedly extends the timescale of subsurface stimulation of prokaryotic processes at the margin site and any other deep, sub-seafloor sediment⁹.

METHODS

Sediment handling, activity and total prokaryotic count measurements. Samples were obtained from Sites 1226 and 1229 on ODP Leg 201 (ref. 29). All sediment was subsampled, aseptically and anaerobically (not molecular genetic samples) at about 4 °C and then rapidly processed for radiotracer or prokaryotic counts or stored for molecular analysis (–80 °C) on board ship. Intact syringe subcores were injected with radiotracer (¹⁴C]bicarbonate or [¹⁴C]acetate, [³⁵S]sulphate, or [³H]thymidine) and incubated at close to temperatures *in situ*; activity was then stopped by freezing or the addition of zinc acetate before processing in the laboratory^{18,30}. Because incubation conditions were not identical to conditions in the original sediment, measured rates might differ from those *in situ*. Total populations were determined by staining with acridine orange coupled with epifluorescence microscopy (AODC, ref. 25); these were conducted mainly on the ship. Total counts were assessed for significant differences with a one-way analysis of variance, using the Tukey–Kramer method for comparing individual means ($\log_{10}$ (minimum significant difference) = 0.59 for $P < 0.05$) and the sum-of-squares simultaneous test procedure for groups of means. Numbers of sulphate-reducing bacteria were estimated from the specific sulphate reduction rate of 0.1 fmol per cell per day (ref. 28) and the measured sulphate reduction rates. This was expressed as a percentage of the AODC total count.

Two types of tracer were used to assess contamination from sea water during drilling: a perfluorocarbon (PFT) and bacteria-sized fluorescent beads (0.5 µm), combined with aseptic handling and subsampling. If present, contamination was concentrated near the core liner. Contamination was low (less than 0.1 µl of sea water per gram of sediment) or below detection in advanced piston cores, all Peru margin samples and to about 300 m BSF at the open ocean site. Contamination was more variable and slightly higher in extended core barrel cores taken below 300 m BSF for the open ocean site (about 0.24 µl of sea water per gram of sediment). A 0.1-µl volume of sea water contains about 50 bacteria, which is 0.000012% of the average prokaryotic population at Site 1229. Because sediment near the core liner was not sampled, and detected contamination was low or absent, subsequent analysis was not subject to seawater contamination.

DNA extraction and polymerase chain reaction (PCR) conditions. DNA was extracted from sediments²⁰ and then stored at –80 °C. PCR was conducted with 16S rRNA gene primers for Bacteria (27F-907R and 27F-1492R), for Archaea (109F-958R), for candidate division JS1 (63F-665R) and primers for *mcrA* genes (ME1 and ME2)^{21,22}. Two primer pairs were also used to amplify *dsrAB* and *dsrA* genes from the sediments, but these did not yield products from any depth despite successful amplifications on other Leg 201 and near-surface sediments.

Cloning and sequencing. PCR products were screened by DGGE to ensure representative amplification as described previously²⁰, and five independent PCR products from each depth (6.7, 30.2, 42.03 and 86.67 m BSF) were pooled and cleaned (Wizard PCR Preps DNA Purification System; Promega). Cloning was conducted with the pGEM-T Easy Vector System (Promega) and inserts were confirmed by PCR with vector-specific M13, 16S rRNA gene or *mcrA* gene primers. Random clones were chosen and sequenced with 27F (Bacteria 16S rRNA genes), 109F (Archaea 16S rRNA genes) and ME1 (*mcrA* genes) primers by using an ABI 3100 Prism Genetic Analyzer (Applied Biosystems). Sequences were checked for chimaeras and assigned to phylogenetic groups by sequence comparison with databases (<http://www.ncbi.nlm.nih.gov/>); assignment was confirmed by phylogenetic tree reconstruction with, first, neighbour-joining with the Jukes and Cantor correction, and second, minimum evolution with the

LogDet/Paralinear distance methods^{21,22}. Sequence lengths used to construct Fig. 2a were as follows: 640–720 base pairs (bp) for Archaea, 315–828 bp for Bacteria (mean 580 bp; 92% more than 450 bp) and 605–720 bp for methanogen *mcrA* sequences.

DGGE analysis. 16S rRNA gene products from Bacteria were reamplified in a nested PCR with primers 357F and 518R; PCR products were analysed by DGGE²⁰. DGGE bands were identified by visually inspecting gels through a mask consisting of 33 horizontal slices, with each slice approximately the width of the brightest band. The 24 DGGE bands obtained were scored (down to the lowest marker band in Fig. 2b) as present (score 1) or absent (score 0) and the data were analysed by principal-component analysis with Minitab Release 14 (Minitab Inc.), using depths as the variables. Other ordination approaches including multi-dimensional scaling, factor analysis and using bands as variables, scoring by the presence or absence of a band and by band density, all provided very similar results, confirming the consistency of our analysis; some cluster analysis approaches showed similar groupings to the ordinations, but others did not. Correlation of the individual principal components with prokaryotic and geochemical variables²⁹ was used to identify the main factors affecting the diversity of Bacteria in depth profiles. The DGGE profile of 88.4 m BSF was excluded from the correlation analysis: inclusion obscured overall relationships because the biodiversity of this depth was so different from that at all other depths.

Received 24 March; accepted 10 May 2005.

- Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
- D'Hondt, S., Rutherford, S. & Spivack, A. J. Metabolic activity of subsurface life in deep-sea sediments. *Science* **295**, 2067–2070 (2002).
- Parkes, R. J. & Wellsbury, P. in *Microbial Diversity and Bioprospecting* (ed. Bull, A. T.) 120–129 (ASM Press, Washington DC, 2004).
- Chapelle, F. H. & Lovley, D. R. Rates of microbial-metabolism in deep coastal-plain aquifers. *Appl. Environ. Microbiol.* **56**, 1865–1874 (1990).
- Lovley, D. R. & Chapelle, F. H. Deep subsurface microbial processes. *Rev. Geophys.* **33**, 365–381 (1995).
- Krumholz, L. R., McKinley, J. P., Ulrich, G. A. & Suflita, J. M. Confined subsurface microbial communities in Cretaceous rock. *Nature* **386**, 64–66 (1997).
- McMahon, P. B., Chapelle, F. H., Falls, W. F. & Bradley, P. M. Role of microbial processes in linking sandstone diagenesis with organic rich clays. *J. Sedim. Petrol.* **62**, 1–10 (1992).
- Wellsbury, P. *et al.* Deep marine biosphere fuelled by increasing organic matter availability during burial and heating. *Nature* **388**, 573–576 (1997).
- Parkes, R. J., Cragg, B. A. & Wellsbury, P. Recent studies on bacterial populations and processes in subseafloor sediments: A review. *Hydrogeol. J.* **8**, 11–28 (2000).
- Teske, A., Wawer, C., Muyzer, G. & Ramsing, N. B. Distribution of sulfate-reducing bacteria in a stratified Fjord (Mariager Fjord, Denmark) as evaluated by most-probable-number counts and denaturing gradient gel electrophoresis of PCR-amplified ribosomal DNA fragments. *Appl. Environ. Microbiol.* **62**, 1405–1415 (1996).
- Coolen, M. J. L., Cypionka, H., Sass, A. M., Sass, H. & Overmann, J. Ongoing modification of Mediterranean Pleistocene sapropels mediated by prokaryotes. *Science* **296**, 2407–2410 (2002).
- Inagaki, F. *et al.* Microbial communities associated with geological horizons in coastal subseafloor sediments from the Sea of Okhotsk. *Appl. Environ. Microbiol.* **69**, 7224–7235 (2003).
- D'Hondt, S. *et al.* Distributions of microbial activities in deep subseafloor sediments. *Science* **306**, 2216–2221 (2004).
- Whelan, J. K., Kanyo, Z., Tarafa, M. & McCaffrey, M. A. Organic matter in Peru Upwelling sediments—analysis by pyrolysis, pyrolysis-gas chromatography, and pyrolysis-gas chromatography mass spectrometry. *Proc. ODP Sci. Results* **112**, 573–587 (1990).
- Mittlerer, R. M. *et al.* Co-generation of hydrogen sulfide and methane in marine carbonate sediments. *Geophys. Res. Lett.* **28**, 3931–3934 (2001).
- Boetius, A. *et al.* A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407**, 623–626 (2000).
- Iversen, N. & Jørgensen, B. B. Anaerobic methane oxidation rates at the sulfate-methane transition in marine sediments from Kattegat and Skagerrak (Denmark). *Limnol. Oceanogr.* **30**, 944–955 (1985).
- Wellsbury, P., Goodman, K., Cragg, B. A. & Parkes, R. J. The geomicrobiology of deep marine sediments from Blake Ridge containing methane hydrate (Sites 994, 995 and 997). *Proc. ODP Sci. Results* **164**, 379–391 (2000).
- Wellsbury, P., Herbert, R. A. & Parkes, R. J. Incorporation of [³H]thymidine by obligate and facultative anaerobic bacteria when grown under defined culture conditions. *FEMS Microbiol. Ecol.* **12**, 87–95 (1993).
- Webster, G., Newberry, C. J., Fry, J. C. & Weightman, A. J. Assessment of bacterial community structure in the deep sub-seafloor biosphere by 16S rDNA-based techniques: a cautionary tale. *J. Microbiol. Methods* **55**, 155–164 (2003).
- Newberry, C. J. *et al.* Diversity of prokaryotes and methanogenesis in deep subsurface sediments from the Nankai Trough, Ocean Drilling Program Leg 190. *Environ. Microbiol.* **6**, 274–287 (2004).

22. Webster, G., Parkes, R. J., Fry, J. C. & Weightman, A. J. Widespread occurrence of a novel division of bacteria identified by 16S rRNA gene sequences originally found in deep marine sediments. *Appl. Environ. Microbiol.* **70**, 5708–5713 (2004).
23. Schippers, A. *et al.* Prokaryotic cells of the deep sub-seafloor biosphere identified as living bacteria. *Nature* **433**, 861–864 (2005).
24. Ibarki, M. Eocene through Pleistocene planktonic Foraminifers off Peru, Leg 112—Biostratigraphy and Paleooceanography. *Proc. ODP Sci Results* **112**, 239–262 (1990).
25. Parkes, R. J., Cragg, B. A., Fry, J. C., Herbert, R. A. & Wimpenny, J. W. T. Bacterial biomass and activity in deep sediment layers from the Peru Margin. *Phil. Trans. R. Soc. Lond. A* **331**, 139–153 (1990).
26. Marchesi, J. R., Weightman, A. J., Cragg, B. A., Parkes, R. J. & Fry, J. C. Methanogen and bacterial diversity and distribution in deep gas hydrate sediments from the Cascadia Margin as revealed by 16S rRNA molecular analysis. *FEMS Microbiol. Ecol.* **34**, 221–228 (2001).
27. Kemp, P. F. & Aller, J. Y. Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiol. Ecol.* **47**, 161–177 (2004).
28. Leloup, J., Quillet, L., Oger, C., Boust, D. & Petit, F. Molecular quantification of sulfate-reducing microorganisms (carrying *dsrAB* genes) by competitive PCR in estuarine sediments. *FEMS Microbiol. Ecol.* **47**, 207–214 (2004).
29. Shipboard Scientific Party, Controls on microbial communities in deeply buried sediments, eastern Equatorial Pacific and Peru Margin sites 1225–1231, 27 January – 29 March 2002. *Proc. ODP Init. Rep.* **201**, 1–81 (2003).
30. Kallmeyer, J., Ferdelman, T. G., Weber, A., Fossing, H. & Jørgensen, B. B. A cold chromium distillation procedure for radiolabeled sulfide applied to sulfate reduction measurements. *Limnol. Oceanogr. Methods* **2**, 171–180 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Leg 201 cruise for assistance in obtaining and processing samples, and T. Daniell for assistance with DNA sequencing. This research used samples and data provided by the ODP. The ODP is sponsored by the US National Science Foundation (NSF) and participating countries under the management of Joint Oceanographic Institutions (JOI), Inc. We thank the European Union and the Natural Environment Research Council (Marine and Freshwater Microbial Biodiversity Programme) for supporting this research financially.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.J.P. (j.parkes@earth.cf.ac.uk).

Male-specific *fruitless* specifies the neural substrates of *Drosophila* courtship behaviour

Devanand S. Manoli^{1,2}, Margit Foss³, Adriana Vilella⁴, Barbara J. Taylor³, Jeffrey C. Hall⁴ & Bruce S. Baker²

Robust innate behaviours are attractive systems for genetically dissecting how environmental cues are perceived and integrated to generate complex behaviours. During courtship, *Drosophila* males engage in a series of innate, stereotyped behaviours that are coordinated by specific sensory cues. However, little is known about the specific neural substrates mediating this complex behavioural programme¹. Genetic, developmental and behavioural studies have shown that the *fruitless* (*fru*) gene encodes a set of male-specific transcription factors (*Fru^M*) that act to establish the potential for courtship in *Drosophila*². *Fru^M* proteins are expressed in ~2% of central nervous system neurons, at least one subset of which coordinates the component behaviours of courtship^{3,4}. Here we have inserted the yeast *GAL4* gene into the *fru* locus by homologous recombination and show that (1) *Fru^M* is expressed in subsets of all peripheral sensory systems previously implicated in courtship, (2) inhibition of *Fru^M* function in olfactory system components reduces olfactory-dependent changes in courtship behaviour, (3) transient inactivation of all *Fru^M*-expressing neurons abolishes courtship behaviour, with no other gross changes in general behaviour, and (4) 'masculinization' of *Fru^M*-expressing neurons in females is largely sufficient to confer male courtship behaviour. Together, these data demonstrate that *Fru^M* proteins specify the neural substrates of male courtship.

The *Fru^M* proteins are generated from sex-specifically spliced transcripts from the P1-*fru* promoter^{2,5} (Fig. 1a, b). Using homologous recombination, we introduced the yeast *GAL4* coding sequence, including start and stop codons, into the *fru^M* coding sequence⁶ (Fig. 1b) and simultaneously deleted the first two codons (ATGATG) of the *fru^M* open reading frame to prevent its translation. Proper integration into *fru* was verified using genomic polymerase chain reaction (PCR). This modified *fru* gene, *fruP1-GAL4*, is null for P1-*fru* function. Staining the central nervous system (CNS) of *fruP1-GAL4* homozygotes revealed no *Fru^M* protein (data not shown). These homozygotes do not show courtship behaviour but appear otherwise normal (Supplementary Fig. S1).

To determine whether *fruP1-GAL4* accurately reflects P1-*fru* expression, we compared the CNS expression patterns of *Fru^M* and a nuclear green fluorescent protein (GFP) marker (UAS-*GFPnls*) driven by *fruP1-GAL4*. Approximately 48 h after puparium formation, when *Fru^M* expression is maximal (~1,500–1,700 cells)³, GFP and *Fru^M* signals are coincident (Fig. 1c, d). The number of *Fru^M*-expressing cells declines to ~1,200–1,300 cells in pharate adults, and remains relatively constant into young adulthood³ (Supplementary Fig. S2). Whether this decrease reflects cell death or transient *Fru^M* expression is unknown. We also compared *Fru^M* expression and *fruP1-GAL4*-driven expression of GFP at later times (72–84 h after puparium formation), as GFP should remain in cells that transiently expressed *fruP1-GAL4*. We simultaneously drove the

expression of UAS-*GFPnls* and UAS-*mCD8GFP*, which encodes a relatively stable membrane-bound form of GFP. Comparison of GFP and *Fru^M* signals revealed that most cells stained positively for GFP at the membrane, and for both *Fru^M* staining and *GFPnls* signal in the nucleus. In ~10% of cells there was neither *Fru^M* staining nor nuclear GFP, but GFP was present at the cell membrane (arrowheads in Fig. 2a; Supplementary Fig. S2), suggesting that in these neurons P1-*fru* expression was transient and the nuclear GFP and *Fru^M* proteins were depleted by turnover, while the more stable mCD8GFP persisted.

The site of *GAL4* insertion in *fruP1-GAL4* is common to P1-derived transcripts in both sexes, allowing us to determine sex-specific differences in the principal features of neurons expressing these transcripts. mCD8GFP expression driven by *fruP1-GAL4* revealed a complex pattern of neuronal projections with many prominently labelled nerve bundles and neuropil structures (Fig. 2b, c). No marked differences were seen between the principal features of the projections of P1-*fru* neurons in males and females, suggesting that *Fru^M* proteins do not specify distinct neural structures or function at the level of pathfinding and early development in the neurons in which they are expressed, but more likely specify their fine connectivity and/or physiology.

We next examined the expression of *fruP1-GAL4* throughout the body to determine whether technical limitations had previously prevented detection of *Fru^M* in other tissues. In all peripheral sensory systems implicated in courtship, we found substantial *fruP1-GAL4* expression in subsets of sensory neurons, but not their associated, non-neuronal support cells (Fig. 3 and Supplementary Fig. S3). *fruP1-GAL4* is expressed in ~100–150 olfactory receptor neurons (ORNs) in each antenna. On the basis of their distribution and CNS glomerular projection patterns (see below), these neurons are mostly from trichoid sensilla, which have been implicated in pheromone detection in other species⁷ (arrow in Fig. 3a). *fruP1-GAL4* is also expressed in about four olfactory receptor neurons within each maxillary palp (Fig. 3c, inset). In the auditory system, *fruP1-GAL4* is expressed in most, if not all, neurons in Johnston's organ, a chordotonal organ found in the second antennal segment⁸ (arrowhead in Fig. 3a), as well as in two small chordotonal organs at the base of the wing (Fig. 3e). This is consistent with the observation that proprioceptive feedback is necessary for proper courtship song^{9,10}. The taste (gustatory) neurons of *Drosophila* innervate sensory bristles on the legs, proboscis and the oral tract¹¹, and *fruP1-GAL4* is expressed in ~20–23 gustatory neurons in the foreleg (Fig. 3f) as well as in ~20–30 gustatory neurons in the proboscis (Fig. 3c). In the visual system, we detect transient pupal *fruP1-GAL4* expression in the retina. Expression is seen in corresponding regions in the periphery of both sexes (data not shown).

The only mechanosensory neurons in which we detect *fruP1-GAL4* expression are the neurons innervating (1) the sex comb bristles on

¹Neurosciences Program and ²Department of Biological Sciences, Stanford University, Stanford, California 94305, USA. ³Department of Zoology, Oregon State University, Corvallis, Oregon 97331-2914, USA. ⁴Department of Biology, Brandeis University, Waltham, Massachusetts 02254, USA.

the male foreleg (Fig. 3f, inset, and Supplementary Fig. S3), (2) the genital clasper bristles, (3) the genital lateral plate bristles, (4) bristles on the ventral analia and (5) the hypandrial bristles associated with the penis apparatus (Fig. 3i and Supplementary Fig. S3). Notably, these are the only places where male-specific morphological specializations of mechanosensory bristles are found. Sex combs are used in grasping the female and spreading her wings during copulation in other species, although their function in *D. melanogaster* is unknown¹². Mechanosensory information transduced through genital claspers and genital lateral plates bristles mediates species-specificity and positioning of the genitalia during attempted copulation¹³. Hypandrial bristles may be involved in the detection of sensory cues that elicit the sequential transfer of seminal fluids and sperm¹⁴.

To determine whether peripheral *fruP1-GAL4* expression represented ectopic GAL4 expression, as has been found with *fru* transgenes¹⁵, we used antibodies against Fru^M and *in situ* hybridization to *fru* transcripts to re-examine peripheral *fru* expression in males and females. We found Fru^M protein and *fru* transcript expression in peripheral neurons, consistent with the *fruP1-GAL4* expression pattern (Fig. 3 and Supplementary Fig. S3).

That Fru^M is expressed in subsets of sensory neurons suggests that males and females may detect distinct sensory stimuli at the level of sensory neurons themselves, or that they might process and perceive such sensory information in different ways. Moreover, these findings strongly suggest that sexual sensory cues are initially recognized in the Fru^M-expressing sensory neurons, and thus that these neurons are entry points for following the flow of specific visual, gustatory, olfactory, auditory and tactile information governing courtship.

We also examined whether Fru^M was expressed in higher-order visual and olfactory neurons. We found limited Fru^M expression in optic lobes³, and *fruP1-GAL4* expression in medullary neurons as well as 4–5 clusters of neurons in the lobula, regions where integration and processing of visual information occurs (Supplementary Fig. S4). In addition, using a UAS-*synaptotagmin-HA* (UAS-*synt-HA*) marker to label presynaptic termini, *fruP1-GAL4* expression is seen in distinct tracts leaving the lobulae, including a major tract projecting to the anterior optical tubercle and superior medial protocerebrum (Supplementary Fig. S4).

The axons of olfactory receptor neurons terminate in antennal lobe glomeruli. *fruP1-GAL4*-directed reporter expression showed processes of *fruP1-GAL4* olfactory receptor neurons projecting primarily to 3–4 glomeruli (DA1, VA1l, VA1m and VL2), with much weaker labelling of other glomeruli (Fig. 2d and Supplementary Fig. S4). We observed dendritic projections to these glomeruli from *fruP1-GAL4* labelled projection neurons adjacent to the antennal lobes (Fig. 2d and Supplementary Fig. S4). Notably, others have shown that the DA1 glomerulus is sexually dimorphic in Hawaiian Drosophilids, and to a lesser extent in *D. melanogaster*¹⁶.

Naive male *Drosophila* typically court other males upon first encountering them, but then sustainably habituate to all males¹⁷. To determine whether Fru^M function in primary and/or secondary olfactory neurons was involved in male–male habituation, we analysed males in which Fru^M was inhibited in the majority of olfactory receptor neurons. This inhibition was achieved by expression of an RNA-mediated interference transgene (UAS-*fru^MIR*) targeting the male-specific amino terminus of Fru^M isoforms⁴. Inhibition of Fru^M in most olfactory receptor neurons (through the *Or83b-GAL4*

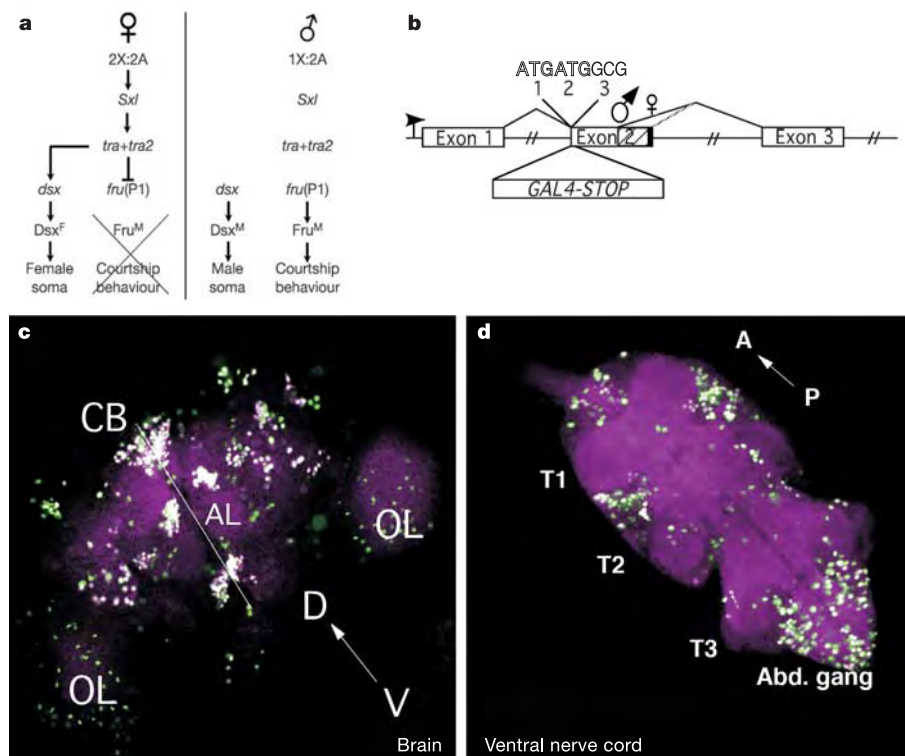


Figure 1 | Male-specific fruitless regulates courtship. **a**, In male flies, the absence of *Sex lethal* (*Sxl*) and *transformer* (*tra*) activity results in the default splicing of P1-*fru* transcripts to produce male-specific isoforms (Fru^M) that are required for courtship behaviour. **b**, The generation of *fruP1-GAL4*. A diagram of the *fru* locus indicates the insertion point of the yeast *GAL4* transcription factor into the P1-*fru* open reading frame by homologous recombination. The arrowhead shows the P1 transcriptional start site. Male and female splice sites are indicated, and the Tra/Tra-2 binding region is

shown in black. Codons 1 and 2 (outline) were deleted upon recombination. **c**, **d**, *fruP1-GAL4*-directed expression accurately reproduces endogenous Fru^M expression patterns. *fruP1-GAL4*-driven nuclear GFP (green) and endogenous Fru^M (magenta) expression in the anterior brain (**c**) and ventral nerve cord (**d**) of a male two-day-old pupa coincide (white) throughout the CNS. Abbreviations used: A, anterior; Abd. gang., abdominal ganglion; AL, antennal lobes (line shows the midline); CB, central brain; D, dorsal; OL, optic lobes; P, posterior; T1–T3, thoracic segments 1–3; V, ventral.

driver), or neurons projecting to the glomeruli labelled by *fruP1-GAL4* (through the *SG18.1-GAL4* driver), resulted in sustained male–male courtship after 1 h of pairing, whereas males expressing a control UAS-*GFP* transgene typically showed a decrease in courtship levels^{18,19} (Fig. 4a). Thus, *Fru^M* function in olfactory receptor neurons and/or secondary olfactory neurons is required for male–male habituation.

As second-order olfactory projection neurons project to the mushroom bodies, we looked for expression of *fruP1-GAL4* in mushroom bodies. Anti-*Fru^M* staining is not seen in pupal mushroom bodies, but weak *Fru^M* staining has been seen in adults in the region of Kenyon cell nuclei^{3,15}. Examining *fruP1-GAL4*-driven UAS-*mCD8GFP* expression in adult flies revealed substantial expression in mushroom body γ -neurons (arrows in Fig. 2d), and in a small number of α/β -neurons (arrowheads) that appeared ~24 h after eclosion, when sexual maturity is attained (Fig. 2d and Supplementary Fig. S4).

Male mushroom body γ -lobes, although not necessary for courtship itself, are necessary for courtship conditioning to mated females (that is, males learn not to court recently mated females, which display high levels of rejection²⁰; J. M. Dura, personal communication). To determine whether *Fru^M* function in mushroom body neurons was necessary for such conditioning, we analysed conditioning in males in which *Fru^M* expression was inhibited in sets of mushroom body neurons by UAS-*fru^MIR* expression. Inhibition of *Fru^M* expression throughout the mushroom bodies (using an *OK107-GAL4* driver) and in γ -neurons (using *H24-GAL4* and *201y-GAL4* drivers) reduced the conditioning response. Restricting

the expression of interfering RNAs to only α/β -neurons (using the *17D-GAL4* driver) had less of an effect (Fig. 4b). Thus, *Fru^M* functions in mushroom bodies to regulate courtship conditioning to mated females. The large number of *Fru^M*-expressing neurons in the mushroom bodies suggests that a significant fraction of the mushroom bodies might function in a manner that is at least in part sex-specific.

There is only minimal *fruP1-GAL4* expression in 'higher-order' centres such as the central complex and much of the proto- and deutocerebrum, structures previously implicated in the generation and coordination of general motor programmes and behaviours in insects²¹ (Fig. 2d and Supplementary Fig. S4). This suggests that *Fru^M* neurons are unlikely to be involved in general processing and coordination of behaviour (see below). *fruP1-GAL4* expression is also not detected in most motor neurons in the ventral nerve cord. This again suggests that *Fru^M*-expressing neurons might modulate, rather than directly mediate, behavioural output (data not shown). One example of such courtship-specific control of conserved neural modules is the generation of song, as the same motor neurons that drive flight also generate courtship song⁹. However, *Fru^M*-expressing neurons might directly control certain outputs of courtship behaviour; for example, *Fru^M*-expressing motor neurons innervate the male-specific muscle of Lawrence, and about eight serotonin-containing, *Fru^M*-expressing neurons provide the sole innervation to some male internal genital organs^{5,15,22,23}. Thus *Fru^M*-expressing neurons might directly mediate output through male-specific structures, and indirectly modulate output dependent on structures common to both sexes.

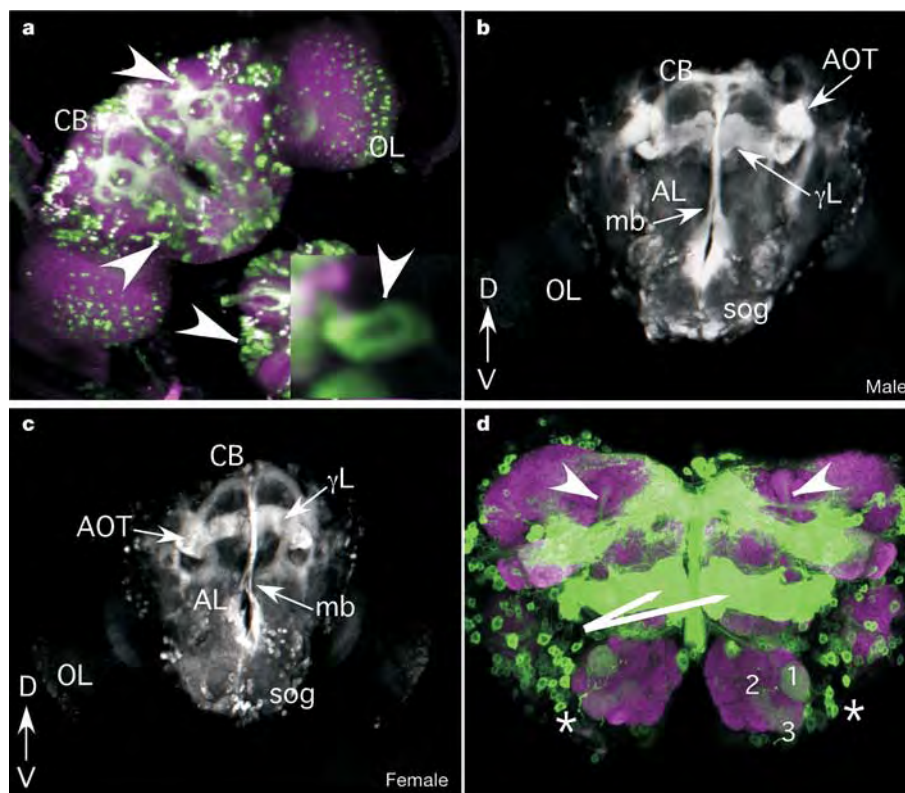


Figure 2 | *fruP1-GAL4* expression in the central nervous system. **a**, *fruP1-GAL4*-driven expression of membrane and nuclear GFP (green) and *Fru^M* (magenta) in pharate adults reveals a limited number of neurons showing membrane GFP expression but neither nuclear GFP nor *Fru^M* staining. This suggests that nuclear GFP and *Fru^M* may be degraded in these cells (arrowheads and inset). **b**, **c**, There are no major differences in *fruP1-GAL4*-driven expression of membrane-bound GFP between males (**b**) and females (**c**), suggesting that *Fru^M* does not specify basic neuronal structures or tracts. Abbreviations used: AOT, anterior optic tubercle;

γ L, mushroom body γ -lobes; mb, median bundle; sog, suboesophageal ganglion (additional abbreviations provided in Fig. 1 legend).

d, *fruP1-GAL4* is expressed in the olfactory system, including in projections from olfactory receptor neurons to antennal lobe glomeruli DA1 (1), VA11 (2), VA1m (3) and VL2 (not shown), projection neurons innervating these glomeruli (asterisks), and mushroom body γ - (arrows) and α/β -neurons (arrowheads). Membrane GFP is shown in green, and neuropil (nc82) staining in magenta.

To determine whether the function of Fru^M -expressing neurons during courtship is necessary, we used *fruP1-GAL4*-directed expression of a temperature-sensitive dynamin allele (*shi^{TS}*) to transiently inactivate these neurons. Transient inactivation of Fru^M -expressing neurons in males at restrictive temperature (31 °C) abolishes courtship behaviour (Fig. 4c; $n = 20$), but grooming, walking and flight behaviours are normal (Supplementary Video S1), suggesting that Fru^M -expressing neurons are largely dedicated to courtship.

We asked whether expression of Fru^M in these neurons is both necessary and sufficient to confer the potential for male courtship by using *fruP1-GAL4*-driven expression of UAS-*tra2IR* to inhibit transformer-2 (*Tra-2*) expression and thus masculinize just the Fru^M -expressing neurons in a female^{5,24} (see Fig. 1a). Strikingly, *fruP1-GAL4/UAS-tra2IR* masculinized females all (10/10) displayed the initial stages of courtship behaviour—orientation and tapping—when paired with a wild-type virgin female (Fig. 4d), but wing and proboscis extension and attempted copulation were not seen. When paired with a wild-type male, these masculinized females were always courted, but showed male-like rejection behaviours, including wing flicking and kicking, and never showed the female rejection response of ovipositor extrusion seen in control females (Fig. 4d).

Similarly, *fruP1-GAL4*-directed expression of individual Fru^M isoforms (as UAS-*fru* or UAS-*fru^M* constructs) in females also conferred certain aspects of courtship behaviour (Fig. 4d). However, the lower level and extent of courtship behaviours in these females suggest that each isoform functions in a non-redundant manner.

We wondered whether such masculinized females might have the potential for more aspects of male courtship than they displayed. As hearing male song is sufficient to induce courtship behaviour in

wild-type males²⁵, we placed multiple *fruP1-GAL4/UAS-tra2IR* masculinized females with a single wild-type male. Indeed, in 10 out of 13 groups containing three *fruP1-GAL4/UAS-tra2IR* females and one wild-type male, male singing was sufficient to elicit wing extension and vibration as well as occasional proboscis extension in a masculinized female that was not being courted (Fig. 4d and Supplementary Fig. S5). No attempts at copulation were observed, perhaps owing to the anatomical restrictions of a female abdomen. Thus *fruP1-GAL4* masculinized females have the potential for more male courtship behaviour than they display when with a single female. This could be because the masculinization/transformation by UAS-*tra2IR* was incomplete or because male identity in tissues other than *fruP1*-expressing neurons is necessary for proper stimulation. The observation that Fru^M function in a distinct subset of neurons is both necessary and largely sufficient to confer the potential for courtship strongly supports the idea that the circuitry underlying innate behaviours might be controlled by dedicated genetic programmes².

Our findings offer new insights into the neuronal circuitry underlying complex behavioural programmes. The existence of Fru^M expression in subsets of all peripheral sensory systems implicated in courtship, as well as second- and third-order neurons in the two sensory systems examined, suggests that specific parts of sensory systems mediate the detection and initial processing of sensory cues relevant to courtship. The lack of overt sexual dimorphism in Fru^M -expressing neurons suggests that Fru^M proteins function to alter fine neuronal connectivity and/or physiology in order to process and transmit information relevant to courtship arousal. That Fru^M -expressing neurons have little (if any) role in other behaviours suggests that these neurons modulate conserved elements

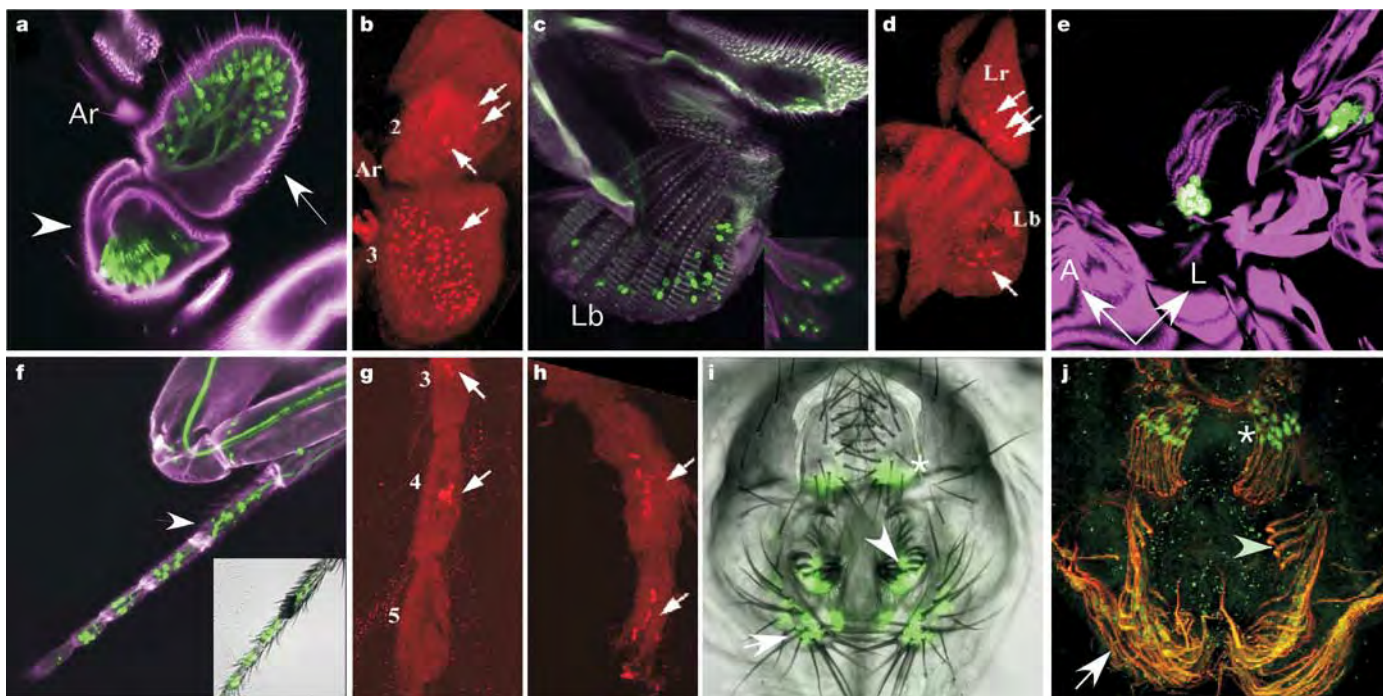


Figure 3 | *fruP1-GAL4* reveals Fru^M expression in regions of the peripheral nervous system implicated in courtship behaviours. Shown are *fruP1-GAL4*-expressing neurons (membrane GFP, green) and autofluorescence (magenta/grey; **a**, **c**, **e**, **f**, **i**) in peripheral nervous system structures. Endogenous Fru^M is found in these locations (arrows in **b**, **d**, **g**, **h**, **j**). **a**, **b**, In the antenna, *fruP1-GAL4* labels 100–150 olfactory sensory neurons in the third antennal segment (arrow in **a**) and auditory neurons of Johnston's organ in the second segment (arrowhead in **a**; Ar, arista). **c**, **d**, In the proboscis, 20–30 gustatory neurons express *fruP1-GAL4*, and 4 olfactory neurons in the maxillary palps are labelled (inset). Lb, labellum; Lr, labrum.

e, In the wing joint, *fruP1-GAL4* labels two clusters of proprioceptive neurons (A, anterior; L, lateral). **f–h**, In the prothoracic leg, *fruP1-GAL4* labels gustatory neurons and mechanosensory neurons associated with the sex combs (arrow in **f**, inset shows brightfield image of leg and sex comb; proximal tarsus segments numbered in **g**; distal tarsus shown in **h**). **i**, **j**, In the male external genitalia, *fruP1-GAL4* labels distinct clusters of mechanosensory neurons associated with bristles on the lateral plates (arrow), the claspers (arrowhead), and the ventral-most part of the analia (asterisk in **i**, **j**), neuronal projections (22C10) are shown in red (**j**).

of the nervous system for courtship-specific behavioural output. Thus, the specification of distinct circuitry for complex innate behavioural programmes might involve the establishment of elements that (1) discriminate specific stimuli from background, (2) integrate such information from multiple sensory modalities, and (3) relay ethologically relevant input to and output from conserved components of the nervous system to generate specific behavioural states, as well as elements that coordinate distinct behavioural modules⁴. A precedent for such a circuit involved in mating behaviour, in which sensory cues detected through male-specific neurons

mediate the coordination of centrally generated behaviours, is seen in nematodes²⁶.

We can now begin to characterize the molecular and cellular processes regulated by *Fru^M* proteins, and examine how these processes act during development to build the potential for male sexual behaviour. Understanding the apparently subtle but nevertheless critical function of *Fru^M* as a transcription factor might help to elucidate the evolutionary strategies through which behavioural programmes are built from or into general components of the nervous system²⁷. We can now also address how specific neurons function to detect or transmit behaviourally relevant sensory cues, integrate this information to perceive the external environment, and process such information to generate and modulate meaningful behavioural output.

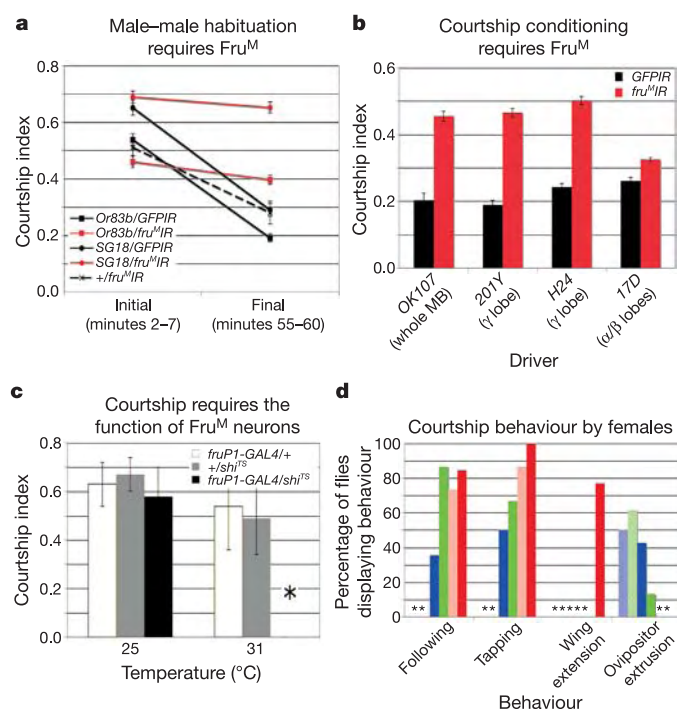


Figure 4 | Function of *Fru^M* neurons in courtship. **a**, Inhibition of *fru^M* expression in primary and/or secondary olfactory neurons reduces male-male habituation. Shown are courtship index (CI) values for pairs of males with *fru^M* inhibition by UAS-*fru^M*IR expression (SG18, $n = 20$; OR83b, $n = 12$), control males expressing UAS-GFPiR ($n = 10$ for SG18 and OR83b drivers) or with UAS-*fru^M*IR alone ($n = 10$). Males showed persistent male-male courtship after the habituation period (SG18, $F_{1,18} = 114.7$; OR83b, $F_{1,18} = 87.6$; $P < 0.001$) in *fru^M*IR but not control animals. **b**, Inhibition of *fru^M* expression in mushroom bodies reduces courtship conditioning in response to mated females. Shown are CI values for males with virgin wild-type females after exposure to a mated wild-type female ($n = 10$ for all groups). RNAi effects, $F_{1,72} = 459.7$; driver effects, $F_{3,72} = 12$; interaction, $F_{3,72} = 30.5$, $P < 0.001$. Homogeneity groups between lines for each treatment: GFPiR, all lines; *fru^M*IR, OK107/201Y, H24, 17D. **c**, Inhibition of synaptic transmission in *fruP1-GAL4*-expressing neurons in males abolishes courtship. Shown are CI values for *fruP1-GAL4/+* ($n = 10$), UAS-*shi^{TS}* ($n = 10$) and *fruP1-GAL4/UAS-shi^{TS}* ($n = 20$) males at permissive (25°C) and restrictive (31°C) temperatures. Following a burst of wing extension (14/20 males, 62 ± 7 s), *fruP1-GAL4/UAS-shi^{TS}* males thereafter displayed no courtship. **d**, Expression of *Fru^M* in and masculinization of *fruP1-GAL4*-expressing neurons in females confers components of courtship behaviour. Females expressing *Fru^M* zinc-finger isoforms A or C show following and tapping behaviour towards a virgin CS male, and decreased levels of ovipositor extrusion when placed with a CS male. Only females masculinized in *fruP1-GAL4*-expressing neurons show wing and sometimes proboscis extension when grouped and placed with a CS male. UAS-transgenes used: sex-common isoforms are light green (*fru^C*; $n = 10$) and light blue (*fru^A*; $n = 13$); male-specific isoforms are green (*fru^{MC}*; $n = 14$), blue (*fru^{MA}*; $n = 15$) and pink (*tra2IR*; $n = 15$). Red bars represent groups containing 1 CS male and 3 *tra2IR* females ($n = 13$ groups). Asterisks (**c**, **d**) indicate no behaviour observed. All error bars indicate s.e.m.

METHODS

Drosophila stocks and culture. The *fruP1-GAL4* line was generated as described below. The UAS-mCD8GFP, UAS-*traF* and UAS-*tra2IR* lines were obtained from the Bloomington Drosophila Stock Center. The Stinger 5 nuclear GFP (UAS-GFPnls) line was a gift from S. Barolo. The UAS-*fru* lines were a gift from S. Goodwin²⁸. The UAS-*shi^{TS}* line was provided by T. Kitamoto²⁹. The UAS-GFPiR line (RNA inhibitory to GFP) was a gift from the Krasnow laboratory. The UAS-*fru^M*IR line has been previously described⁴. All stocks and crosses were maintained at 25°C except for those using UAS-*shi^{TS}*, UAS-*tra2IR* and UAS-*fru^M*IR flies, for which crosses were performed at 18°C, 29°C and 29°C, respectively.

Generation of *fruP1-GAL4* through homologous recombination. The techniques for homologous recombination were adapted from previous studies⁵. Fragments containing ~3 kb of sequence 5' and 3' to the *fru^M* start codon were independently cloned. The first three codons of the GAL4 coding sequence were added to the 3' end of the 5' fragment, with codons 2 and 3 of GAL4 altered to create a *HindIII* site, and a *SacII* site was added to the 5' end of the fragment. The 3' fragment began with codon 3 of the *fru^M* coding sequence (the first 2 codons were deleted), and was flanked on the 5' end by a *BamHI* site and on the 3' end by a *StuI* site. The GAL4 coding sequence was amplified using primers with mutations to change codons 2 and 3, and included a *BamHI* site after the stop codon. Fragments were ligated into the pWhiteOut2 P-element transformation vector (a gift from J. Sekelsky) and transformants were generated using standard techniques.

After transformation, multiple lines containing the donor element (pWhite-Out2 construct) were crossed to a UAS-mCD8GFP line to verify absence of ectopic GAL4 expression. Donor lines were then crossed to obtain progeny that contained the donor elements as well as heat-shock inducible FLPase and I-Sce. Larvae were heat shocked for 1 h on days 3 and 4. Individual progeny containing all three elements were then crossed to a UAS-mCD8GFP line and progeny were examined for GFP expression, indicating mobilization of the donor element, splicing and expression of GAL4. Approximately 1,500 individual crosses were screened and eight independent insertion events were isolated and confirmed using genomic PCR. These lines were then crossed to a nuclear GFP reporter, and co-expression in *Fru^M*-expressing neurons in the CNS was verified by immunohistochemistry using standard techniques⁴.

Tissue dissection, staining and imaging. CNS and peripheral tissue were dissected and fixed using standard techniques⁴. Additional *fruP1-GAL4*-expressing neurons were seen in specific peripheral locations with two copies of the reporter transgene. Analysis presented is from animals with one reporter.

Rat anti-*Fru^M* antibody was used at 1:300, rat anti-HA (Roche) was used at 1:100, mouse monoclonal nc82 was used at 1:20, and Cy3-conjugated goat anti-rat and goat anti-mouse antibodies were used at 1:1,000 (Jackson Immuno-research). For colorimetrically-visualized tissue, flies were cryosectioned and visualized as described³⁰, but were labelled with anti-*Fru^M* antibody (1:300) and an alkaline-phosphatase-conjugated goat anti-rat secondary antibody (1:200). For the whole mounts, fixed tissue was incubated for 5 min in PBS with 5% Triton X-100, rinsed and processed using anti-*Fru^M* antibody (1:300) and goat anti-rat AlexaFluor555-conjugated secondary antibody (Molecular Probes/Invitrogen). The samples were mounted in Vectashield mounting media (Vector Labs) and imaged using a BioRad MRC 1024 microscope, or mounted in ProLong reagent (Molecular Probes), for antibody and *in situ* hybridization preparations of peripheral tissue, and imaged on a Zeiss LSM510 Meta scanning confocal microscope.

In situ hybridization on 20-μm tissue sections was performed using the previously described S1 riboprobe³⁰.

Behavioural assays. Courtship assays were performed at ZT (circadian time) 6–10 h with males entrained in isolation for 3–5 days in 12 h light/dark cycles, and 3–5-day-old virgin females; assays were performed at 25 °C except as noted below⁴. Courtship index (CI) was calculated as the percentage of time spent courting (including following, tapping, wing and proboscis extension and attempted/successful copulation) divided by the total observation time. For habituation assays, sibling males were paired for 1 h and the courtship index was calculated for minutes 2–7 and 55–60. For courtship conditioning assays, males were paired in a mating chamber with a mated CS (Canton-S) female for 45–60 min, and then placed into a new chamber with a virgin CS female. For experiments using *UAS-shi^{TS}* flies, crosses were performed and the flies were raised in isolation for 6–10 days after eclosion at 18 °C, entrained at 25 °C for two days (as above) and then assayed at 25 °C and 31 °C. Isolated animals were warmed for 10–15 min at 31 °C before courtship assays.

Statistical analysis. For comparisons of male habituation, final values of CI for males expressing *fru^{MIR}* or *GFPIR* were compared using a one-way analysis of variance (ANOVA). As the driver lines did not have a common genetic background, lines were analysed independently to determine whether changes in final CIs were significant. For comparison of mushroom-body-mediated effects on courtship conditioning, a two-way ANOVA showed a significant effect for both *GAL4* lines and *fru^{MIR}* expression (see Fig. 4 legend). Tukey and Bonferroni post-tests were used to determine homogeneity between drivers for each treatment.

Received 8 April; accepted 1 June 2005.

Published online 15 June 2005.

- Greenspan, R. J. & Ferveur, J. F. Courtship in *Drosophila*. *Annu. Rev. Genet.* **34**, 205–232 (2000).
- Baker, B. S., Taylor, B. J. & Hall, J. C. Are complex behaviors specified by dedicated regulatory genes? Reasoning from *Drosophila*. *Cell* **105**, 13–24 (2001).
- Lee, G. et al. Spatial, temporal, and sexually dimorphic expression patterns of the *fruitless* gene in the *Drosophila* central nervous system. *J. Neurobiol.* **43**, 404–426 (2000).
- Manoli, D. S. & Baker, B. S. Median bundle neurons coordinate behaviours during *Drosophila* male courtship. *Nature* **430**, 564–569 (2004).
- Ryner, L. C. et al. Control of male sexual behavior and sexual orientation in *Drosophila* by the *fruitless* gene. *Cell* **87**, 1079–1089 (1996).
- Gong, W. J. & Golic, K. G. Ends-out, or replacement, gene targeting in *Drosophila*. *Proc. Natl Acad. Sci. USA* **100**, 2556–2561 (2003).
- Keil, T. A. Fine structure of the pheromone-sensitive sensilla on the antenna of the hawkmoth, *Manduca sexta*. *Tissue Cell* **21**, 139–151 (1989).
- Boekhoff-Falk, G. Hearing in *Drosophila*: development of Johnston's organ and emerging parallels to vertebrate ear development. *Dev. Dyn.* **232**, 550–558 (2005).
- Ewing, A. W. The neuromuscular basis of courtship song in *Drosophila*: The role of direct and axillary wing muscles. *J. Comp. Physiol.* **130**, 87–93 (1979).
- Smith, S. A. & Shepherd, D. Central afferent projections of proprioceptive sensory neurons in *Drosophila* revealed with the enhancer-trap technique. *J. Comp. Neurol.* **364**, 311–323 (1996).
- Scott, K. C., Taubman, A. D. & Geyer, P. K. Enhancer blocking by the *Drosophila* gypsy insulator depends upon insulator anatomy and enhancer strength. *Genetics* **153**, 787–798 (1999).
- Spieth, H. T. Courtship behavior in *Drosophila*. *Annu. Rev. Entomol.* **19**, 385–405 (1974).
- Acebes, A., Cobb, M. & Ferveur, J. F. Species-specific effects of single sensillum ablation on mating position in *Drosophila*. *J. Exp. Biol.* **206**, 3095–3100 (2003).
- Wolfner, M. F. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. *Heredity* **88**, 85–93 (2002).
- Billeter, J. C. & Goodwin, S. F. Characterization of *Drosophila fruitless-gal4* transgenes reveals expression in male-specific *fruitless* neurons and innervation of male reproductive structures. *J. Comp. Neurol.* **475**, 270–287 (2004).
- Kondoh, Y., Kaneshiro, K. Y., Kimura, K. & Yamamoto, D. Evolution of sexual dimorphism in the olfactory brain of Hawaiian *Drosophila*. *Proc. R. Soc. Lond. B* **270**, 1005–1013 (2003).
- Vaia, L. J., Napolitano, L. M. & Tompkins, L. Identification of stimuli that mediate experience-dependent modification of homosexual courtship in *Drosophila melanogaster*. *Behav. Genet.* **23**, 91–97 (1993).
- Larsson, M. C. et al. *Or83b* encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. *Neuron* **43**, 703–714 (2004).
- Shyamala, B. V. & Chopra, A. *Drosophila melanogaster* chemosensory and muscle development: identification and properties of a novel allele of *scalloped* and of a new locus, SG18.1, in a Gal4 enhancer trap screen. *J. Genet.* **78**, 87–97 (1999).
- McBride, S. M. et al. Mushroom body ablation impairs short-term memory and long-term memory of courtship conditioning in *Drosophila melanogaster*. *Neuron* **24**, 967–977 (1999).
- Strauss, R. The central complex and the genetic dissection of locomotor behaviour. *Curr. Opin. Neurobiol.* **12**, 633–638 (2002).
- Acebes, A., Grosjean, Y., Everaerts, C. & Ferveur, J. F. Cholinergic control of synchronized seminal emissions in *Drosophila*. *Curr. Biol.* **14**, 704–710 (2004).
- Lee, G. & Hall, J. C. Abnormalities of male-specific FRU protein and serotonin expression in the CNS of *fruitless* mutants in *Drosophila*. *J. Neurosci.* **21**, 513–526 (2001).
- Anand, A. et al. Molecular genetic dissection of the sex-specific and vital functions of the *Drosophila melanogaster* sex determination gene *fruitless*. *Genetics* **158**, 1569–1595 (2001).
- Eberl, D. F., Duyk, G. M. & Perrimon, N. A genetic screen for mutations that disrupt an auditory response in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **94**, 14837–14842 (1997).
- Liu, K. S. & Sternberg, P. W. Sensory regulation of male mating behavior in *Caenorhabditis elegans*. *Neuron* **14**, 79–89 (1995).
- Shah, N. M. et al. Visualizing sexual dimorphism in the brain. *Neuron* **43**, 313–319 (2004).
- Song, H. J. et al. The *fruitless* gene is required for the proper formation of axonal tracts in the embryonic central nervous system of *Drosophila*. *Genetics* **162**, 1703–1724 (2002).
- Kitamoto, T. Conditional disruption of synaptic transmission induces male–male courtship behavior in *Drosophila*. *Proc. Natl Acad. Sci. USA* **99**, 13232–13237 (2002).
- Goodwin, S. F. et al. Aberrant splicing and altered spatial expression patterns in *fruitless* mutants of *Drosophila melanogaster*. *Genetics* **154**, 725–745 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank T. Clandinin and members of the Baker laboratory for discussions and comments on this manuscript, J. Sekelsky for the gift of the pWhiteOut2 vector, A. O'Reilly and M. Simon for technical advice, Y.-S. Liu for help with dissections, M. Siegal for help with statistics, and G. Bohm for preparation of culture materials and fly food. This work was supported by an NINDS grant to B.J.T., J.C.H. and B.S.B.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.S.B. (bbaker@pmgm2.stanford.edu).

EphrinB2 is the entry receptor for Nipah virus, an emergent deadly paramyxovirus

Oscar A. Negrete¹, Ernest L. Levroney¹, Hector C. Aguilar¹, Andrea Bertolotti-Ciarlet⁴, Ronen Nazarian¹, Sara Tajyar¹ & Benhur Lee^{1,2,3}

Nipah virus (NiV) is an emergent paramyxovirus that causes fatal encephalitis in up to 70 per cent of infected patients¹, and there is evidence of human-to-human transmission². Endothelial syncytia, comprised of multinucleated giant-endothelial cells, are frequently found in NiV infections, and are mediated by the fusion (F) and attachment (G) envelope glycoproteins. Identification of the receptor for this virus will shed light on the pathobiology of NiV infection, and spur the rational development of effective therapeutics. Here we report that ephrinB2, the membrane-bound ligand for the EphB class of receptor tyrosine kinases (RTKs)³, specifically binds to the attachment (G) glycoprotein of NiV. Soluble Fc-fusion proteins of ephrinB2, but not ephrinB1, effectively block NiV fusion and entry into permissive cell types. Moreover, transfection of ephrinB2 into non-permissive cells renders them permissive for NiV fusion and entry. EphrinB2 is expressed on endothelial cells and neurons^{3,4}, which is consistent with the known cellular tropism for NiV⁵. Significantly, we find that NiV-envelope-mediated infection of microvascular endothelial cells and primary cortical rat neurons is inhibited by soluble ephrinB2, but not by the related ephrinB1 protein. Cumulatively, our data show that ephrinB2 is a functional receptor for NiV.

Emerging viral pathogens present a critical threat to global health and economies. NiV, and the related Hendra virus (HeV), are members of the newly defined Henipavirus genus of the *Paramyxoviridae* (refs 6,7), and are designated as priority pathogens in the National Institute of Allergy and Infectious Diseases' Biodefense Research Agenda. Since 1999, NiV outbreaks have occurred in Malaysia, Singapore and Bangladesh^{1,8}, and have the potential to severely affect the pig-farming industry⁹.

NiV exhibits an unusually broad host range including humans, pigs, dogs, cats, horses, guinea pigs, hamsters and fruit bats (NiV's presumptive natural host)^{6,10,11}. Such a broad range of host tropism is rare among extant paramyxoviruses. With the possible exception of fruit bats, the disease mortality of all other hosts has been established for both natural or experimental infection^{11,12}. However, the mortality rate in pigs is less than 5% even though the transmission rate approaches 100% (refs 6,9), suggesting that zoonotic transmission to humans has increased the pathogenicity of the virus.

Endothelial cells are the major cellular targets for NiV, and hence syncytial endothelial cells in blood vessels are considered a characteristic feature of Nipah viral disease⁵. The fusion (F) and attachment (G) proteins of NiV mediate syncytia formation, and cell lines from many animal species are permissive for NiV-envelope-mediated fusion^{13,14}, suggesting that the receptor for NiV is highly conserved.

To establish whether NiV-G determines the known cell line tropism of NiV, we generated an immunoadhesin by fusing the

ectodomain of NiV-G with the Fc region of human IgG1 (NiV-G-Fc). NiV-G-Fc bound to fusion-permissive 293T, HeLa and Vero cells^{13,15}, but not to non-permissive Chinese hamster ovary (CHO-pgsA745), pig kidney fibroblast (PK13)¹³ and human Raji B cells (Fig. 1a). NiV-G-Fc immunoprecipitated a 48 kDa band from the surfaces of permissive 293T and Vero cells, but not non-permissive CHO-pgsA745 cells (Fig. 1b). Analysis identified a deletion of 28 amino acids (see Supplementary Table 1) in the globular ectodomain of NiV-G, which is produced as an Fc-fusion dimer at wild-type levels, but no longer binds to the surfaces of permissive cells (Fig. 1b). This deletion mutant (Δ 28NiV-G-Fc) was used as a negative control in preparative immunoprecipitation experiments to purify the putative NiV receptor. Parallel portions of the gel containing the 48 kDa band immunoprecipitated by NiV-G-Fc, but not by Δ 28NiV-G-Fc, were analysed by trypsin digestion and mass spectrometry. Only one transmembrane protein was uniquely identified in the NiV-G-Fc sample versus the Δ 28NiV-G-Fc sample. Two independent tryptic fragments of 12 and 17 amino acids each identified the protein as ephrinB2 (Supplementary Fig. 1).

EphrinB2 is essential for vasculogenesis and axonal guidance, and is expressed on endothelial cells, neurons and smooth muscle cells surrounding small arteries and arterioles^{16,17}—an expression pattern

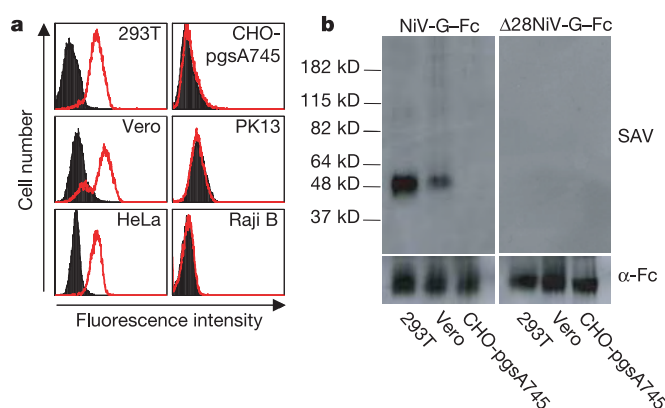


Figure 1 | Soluble NiV-G binds to a 48 kDa membrane protein. **a**, Equal amounts of NiV-G-Fc (thick line) or Fc-only (filled histogram) were incubated with permissive 293T, Vero or HeLa cells, or non-permissive CHO-pgsA745, PK13, or human Raji B cells. Cell-surface binding was detected by a phycoerythrin- (PE-) conjugated anti-human IgG secondary antibody. **b**, Cell surface proteins from permissive 293T and Vero cells, or non-permissive CHO-pgsA745 cells, were biotinylated, immunoprecipitated by NiV-G-Fc or Δ 28NiV-G-Fc, run on a non-denaturing SDS-PAGE gel and detected by western blotting with HRP-conjugated streptavidin (SAV) or anti-human Fc (α -Fc).

¹Department of Microbiology, Immunology and Molecular Genetics, ²Department of Pathology and Laboratory Medicine, ³UCLA AIDS Institute, David Geffen School of Medicine, UCLA, Los Angeles, California 90095, USA. ⁴Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

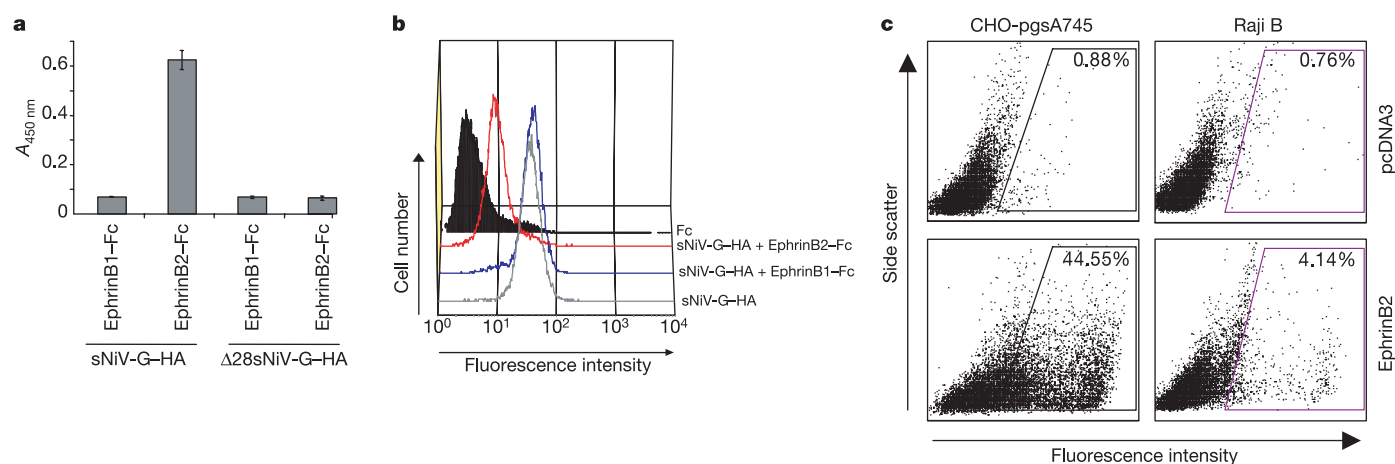


Figure 2 | The ectodomain of NiV-G binds specifically to ephrinB2. **a**, The soluble HA-tagged ectodomain of NiV-G (sNiV-G-HA) bound to ephrinB2-Fc but not ephrinB1-Fc in an ELISA (see Methods). $\Delta 28$ sNiV-G-HA, with an identical deletion as in $\Delta 28$ NiV-G-Fc, did not bind to ephrinB2-Fc or ephrinB1-Fc. $A_{450 \text{ nm}}$, absorbance at 450 nm. One representative experiment out of three is shown. Data are averages of triplicates $\pm$ s.d. **b**, $10 \mu\text{g ml}^{-1}$ of

ephrinB2-Fc but not ephrinB1-Fc was able to block sNiV-G-HA-binding to permissive 293T cells. sNiV-G-HA-binding was detected by a mouse monoclonal anti-HA antibody followed by a PE-conjugated anti-mouse IgG secondary antibody. **c**, NiV-G-Fc bound to ephrinB2-transfected but not to pcDNA3-transfected CHO-pgsA745 and human Raji B cells. Cell-surface binding was detected as in Fig. 1a.

highly concordant with the known cellular tropism of NiV⁵. Using a soluble HA-tagged ectodomain of NiV-G (sNiV-G-HA), we demonstrated that NiV-G bound directly to soluble ephrinB2-Fc, but not to ephrinB1-Fc, in an enzyme-linked immunosorbent assay (ELISA) (Fig. 2a). EphrinB1 is the most closely related ephrin to ephrinB2. Additionally, ephrinB2-Fc, but not ephrinB1-Fc, competed readily for sNiV-G-HA-binding on permissive 293T cells (Fig. 2b), and NiV-G-Fc bound to ephrinB2-transfected, but not to pcDNA3-transfected, CHO-pgsA745 and human Raji B cells (Fig. 2c). Cumulatively, these data demonstrate a direct and specific association between NiV-G and ephrinB2.

Because endothelial syncytia are a hallmark of NiV disease⁵, we

investigated whether ephrinB2 was required for NiV-envelope-mediated syncytia formation. We used a luciferase-reporter-based fusion assay driven by T7-polymerase that has been used extensively to examine viral-envelope-mediated cell-cell fusion^{18,19}. NiV-F/G proteins mediated fusion with permissive 293T or Vero cells, but not with non-permissive PK13 or human Raji B cells (Fig. 3a). No fusion was seen in the absence of NiV-G. Again, soluble ephrinB2, but not ephrinB1, significantly inhibited NiV-F/G-mediated cell-cell fusion (Fig. 3b). Transfection of ephrinB2, but not ephrinB1 or green fluorescent protein (GFP), into human Raji B cells rendered them permissive for NiV-envelope-mediated fusion (Fig. 3c). This fusion was inhibited by soluble ephrinB2 or EphB4 (a cognate

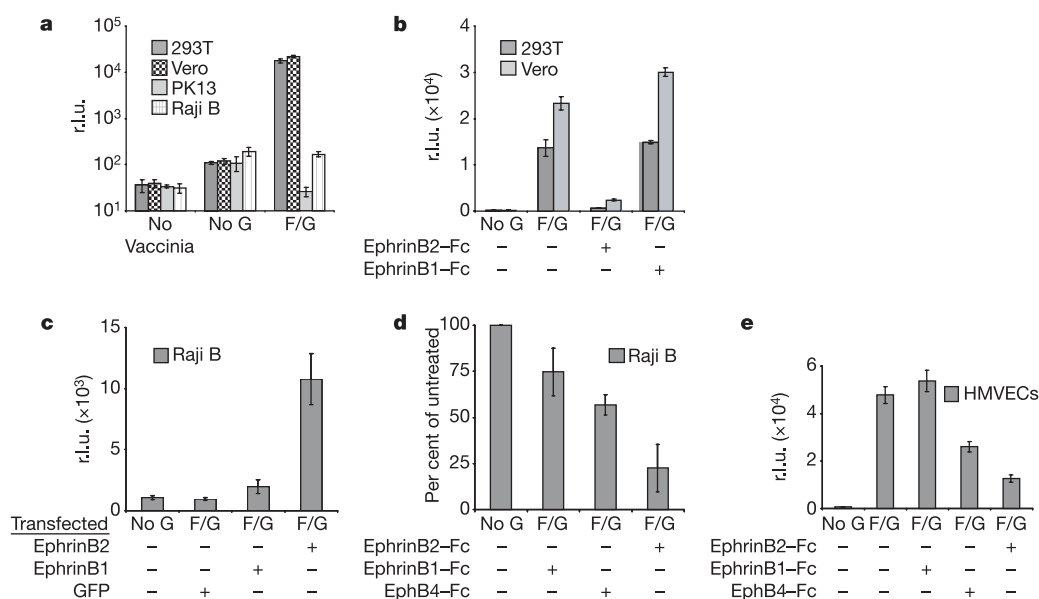


Figure 3 | EphrinB2 is necessary for NiV fusion. **a**, NiV-F/G-expressing 'effector' PK13 cells were placed on permissive (293T or Vero cells) or non-permissive (PK13 or human Raji B) 'target' cells and fusion quantified as described in Methods. **b**, Fusion assay was performed as in **a** for 293T and Vero cells except that ephrinB2-Fc or ephrinB1-Fc ($10 \mu\text{g ml}^{-1}$) was added to the target cells 30 min before addition of NiV-envelope-expressing effector cells. **c**, Fusion assay performed with transfected Raji B target cells

and PK13 effector cells. **d**, Inhibition studies on Raji B cells were performed as in **b** (EphB4-Fc: $100 \mu\text{g ml}^{-1}$). Fusion in each case was normalized to that obtained in the absence of any blocking reagent. **e**, Fusion assay between microvascular endothelial target cells and NiV envelope expressing PK13 effector cells as described. Inhibition studies were performed as in **d**. Error bars indicate $\pm$ s.d. from at least two independent experiments.

receptor for ephrinB2), but not ephrinB1 (Fig. 3d). Significantly, NiV-F/G-expressing cells also fused with human microvascular endothelial cells (HMVECs) in a manner that could be inhibited by soluble ephrinB2 or EphB4, but not ephrinB1 (Fig. 3e). Thus, NiV fusion on cell lines, and on an *in vivo* target cell for NiV infection, is dependent on ephrinB2.

Next we determined whether ephrinB2 could also mediate NiV infection. NiV is a Biosafety Level- (BSL-) 4 pathogen, so we developed a virion-based infection assay that does not require the use of a BSL-4 facility. Heterologous viral envelopes can be pseudotyped onto a recombinant vesicular stomatitis virus (VSV) expressing red fluorescent protein (RFP), but lacking its own envelope (VSV- Δ G-RFP)²⁰. VSV- Δ G-RFP bearing the NiV-F/G proteins was used to infect permissive 293T or Vero cells, resulting in cells expressing RFP (Fig. 4a, b). Viral entry was dependent on NiV-F/G because it was neutralized by NiV-F/G-specific antiserum (Fig. 4a).

VSV-F/G-RFP infection was blocked by ephrinB2-Fc, but not ephrinB1-Fc, while infection by VSV-RFP bearing its own envelope (VSV-G) was not inhibited by either soluble ephrin (Fig. 4b).

Transfection of ephrinB2 into non-permissive CHO-pgsA745 cells rendered them permissive for viral entry (Fig. 4c). CHO-pgsA745 is a mutant CHO cell line that does not express cell surface heparan sulphate proteoglycans²¹. Heparan sulphate has been described as an attachment or entry receptor for many viruses, and may confound the search for bona fide viral receptors that mediate membrane fusion²². Thus, our observation that ephrinB2, in the absence of cell surface heparan sulphates, could mediate viral entry strongly suggests that ephrinB2 is a functional receptor for NiV entry. Finally, NiV-F/G-pseudotyped VSV was also able to infect primary cortical rat neurons and HMVECs (Fig. 4d, e)—two cell types that can be infected *in vivo*⁵. NiV-F/G infection of rat neurons and HMVECs was inhibited by soluble ephrinB2, but not ephrinB1. EphrinB2 inhibited

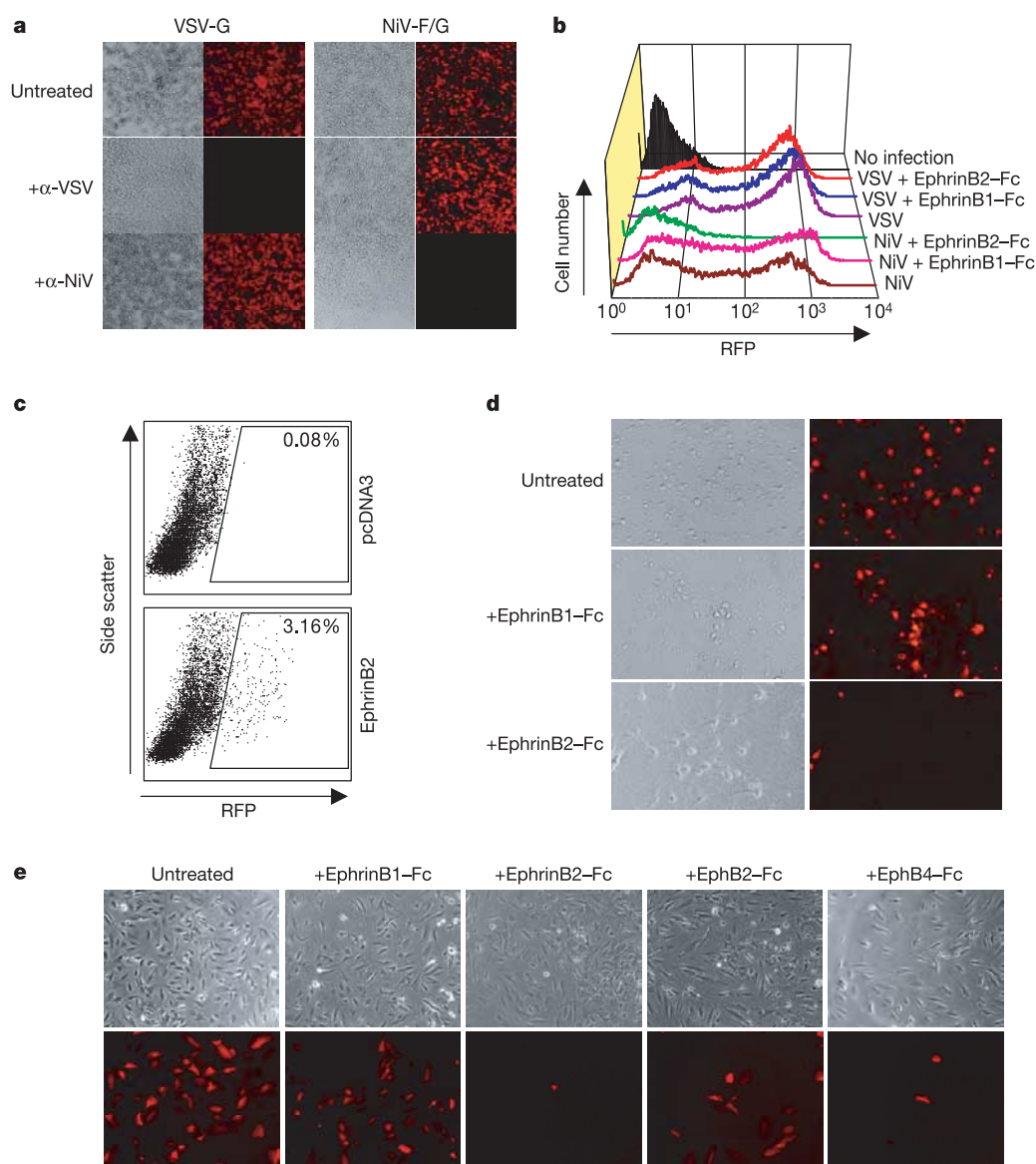


Figure 4 | EphrinB2 mediates entry of NiV-F/G pseudotyped viruses. **a**, VSV-G or NiV-F/G mediated entry into 293T cells was neutralized specifically by their respective antisera. Matched phase-contrast and fluorescent images are shown. **b**, NiV-F/G or VSV-G pseudotyped viruses were used to infect Vero cells in the presence or absence of ephrinB1-Fc or ephrinB2-Fc ($10 \mu\text{g ml}^{-1}$). RFP production was analysed by FACS. **c**, Human ephrinB2- or pcDNA3-transfected CHO-pgsA745 cells were

infected with NiV-F/G pseudotyped VSV-RFP and FACS-analysed for RFP production. **d**, NiV-F/G pseudotyped VSV-RFP viruses were used to infect cortical rat neurons in the presence or absence of ephrinB1-Fc or ephrinB2-Fc ($10 \mu\text{g ml}^{-1}$). Representative matched phase-contrast and fluorescent images are shown. **e**, Additional inhibition studies were performed with HMVECs (ephrinB1/B2-Fc, $10 \mu\text{g ml}^{-1}$; EphB2/B4-Fc, $100 \mu\text{g ml}^{-1}$).

NiV-F/G pseudotype infection of primary rat neurons by 76% compared to ephrinB1 (average number of infected cells per field $\pm$ s.d.: 5.7 ± 4.3 versus 23.5 ± 12.7 for ephrinB2 versus ephrinB1 inhibition, respectively; $P < 0.0001$, Student's *t*-test). Additionally, we show that soluble EphB4 and EphB2 (the cognate receptors for ephrinB2) significantly inhibited NiV-F/G-mediated infection of HMVECs (Fig. 4e). Our use of HMVECs and primary rat neurons to show that NiV-envelope-mediated entry occurred in an ephrinB2-dependent manner strongly suggests that ephrinB2 is also a functional receptor for NiV entry *in vivo*.

In histopathological studies on patients who had succumbed to NiV infection, viral antigen can be detected in unequivocal amounts in relatively few cellular subtypes, such as neurons, endothelial cells and smooth muscle cells surrounding small arteries⁵. This is in concordance with the expression pattern of ephrinB2; in *lacZ* 'knock-in' mice, ephrinB2 was specifically expressed in endothelial cells, neurons and in smooth muscle cells surrounding arterioles^{16,17}. The identification of ephrinB2 as the NiV receptor largely explains the *in vivo* tropism of the virus. EphrinB2 is a critical gene involved in embryogenic development, and has established roles in vasculogenesis and axonal guidance^{3,4}. Ephrin genes are highly conserved and have been found in all animal species examined³. Thus, the conservation of ephrinB2 may also explain the unusually broad tropism of NiV. However, we note that not all cell lines are competent to express ephrinB2 well (our unpublished observations), and we have not conducted exhaustive gain-of-function experiments in all cell lines described to be non-permissive for NiV-envelope-mediated fusion¹³. Therefore, although ephrinB2 is clearly a functional receptor for NiV entry, it is possible that other factors in addition to ephrinB2 expression are required for productive NiV entry and replication.

Both ephrinB2 and its cognate receptor EphB4 have tyrosine signalling and PDZ (Postsynaptic density protein-95, Discs-large, Zonula occludens-1) binding motifs in their cytoplasmic domains²³. 'Forward' signalling mediated by EphB4 facilitates anti-adhesive and repulsive behaviour upon contact with ephrinB2-expressing cells, while ephrinB2 'reverse' signalling facilitates propulsive adhesion upon contact with EphB4-expressing cells. If NiV-G acts like EphB4 and binds to ephrinB2, but lacks the property of reverse signalling, perhaps only forward propulsion will ensue. We speculate that this might act to recruit more endothelial cells to areas of NiV replication. Indeed, signalling-deficient EphB4 on tumour cells can promote invasion by ephrinB2-expressing endothelial cells²⁴. It will be interesting to re-examine pathological specimens for increased angiogenesis in areas of NiV replication. It is also possible that PDZ binding domains, and other proteins known to interact with the cytoplasmic domain of ephrinB2, may play a role in the productive entry of NiV.

HeV appears to have a similar cellular tropism to NiV¹³, although NiV appears to be more pathogenic. Experiments are continuing to determine whether HeV also uses ephrinB2, or ephrinB2-related molecules, as its entry receptor. Discovery of ephrinB2 as the NiV receptor will facilitate screening of small-molecule antagonists to block NiV entry: molecules targeting NiV-G may be potential antivirals, whereas molecules targeting ephrinB2 may have applications in the field of angiogenesis. The recent and repeated outbreaks of NiV in Bangladesh¹ emphasize the importance of the search for vaccines and therapeutics against this emerging pathogen. Identifying the NiV receptor will contribute to these continuing efforts.

METHODS

Cells and reagents. Primary rat cortical neurons were dissected and cultured from embryonic day 17 Sprague-Dawley rats as described²⁵, and plated 2 weeks before infection. HMVECs immortalized with the human telomerase catalytic protein (hTERT)²⁶ were a gift from R. Shao. Soluble Fc-fusion proteins of ephrins and Eph receptors were obtained from R&D Systems. Sequence-verified human ephrinB2 clones were obtained from Origene (CMV-driven clones) and Open Biosystems (T7-driven clones). The open reading frame of

human ephrinB2 was also subcloned into pcDNA3 (Invitrogen) in frame with a C-terminal V5 epitope tag.

Identification of NiV-ephrinB2 interaction. See Supplementary Methods 1 for details of soluble NiV-G production. 293T, Vero or CHO-pgsA745 cells were cell-surface biotinylated using EZ-link Sulfo-NHS-LC-LC-Biotin reagent (Pierce). Each 100-mm dish of cells was lysed (50 mM Tris-HCl, 150 mM NaCl, and 1% Triton X-100, pH 8.0 with protease inhibitors), clarified by centrifugation and pre-cleared by one round of mock immunoprecipitation with Fc-only protein using protein G-coupled magnetic beads (Dyna). Pre-cleared lysates were immunoprecipitated with NiV-G-Fc or Δ 28NiV-G-Fc previously crosslinked to protein G beads (20 mM dimethyl-pimelimidate-HCl in 0.2 M triethanolamine) (Sigma), separated by non-denaturing SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and analysed by western blotting with horseradish peroxidase- (HRP-) conjugated streptavidin or anti-human Fc (Pierce). 293T cells (3×10^7) were used for preparative immunoprecipitations, and proteins were visualized using Silver Stain Plus (BioRad). Parallel portions of the gel containing a specific band immunoprecipitated by NiV-G-Fc, but not Δ 28NiV-G-Fc, were excised, digested in the gel with sequencing-grade trypsin and subjected to peptide sequencing by tandem mass spectrometry (MS/MS). A Finnigan ion trap mass spectrometer LCQ coupled with a high-performance liquid chromatography (HPLC) system running a 75- μ m inner diameter C18 column was used. MS/MS spectra were used to search the most recent non-redundant protein database from GenBank with the ProtQuest software suite (ProtTech).

Fusion assay. Fusion assays were performed essentially as described^{18,19}. Briefly, effector cells (PK13) were transfected with 0.3 μ g of codon-optimized NiV-F and G expression plasmids, 0.6 μ g of T7-luciferase, and 0.8 μ g of pcDNA3 per 6-well plate using Lipofectamine 2000 reagent (Invitrogen). The sequences of the codon-optimized genes have been deposited into GenBank (AY816748 and AY816746 for F and G, respectively, based on the original sequences described for NiV-F and -G in ref. 6.). The DNA amount was always kept constant with pcDNA3. Target cells (293T, Vero, HMVECs) grown in a 24-well plate were infected with vaccinia virus (vTF1.1) expressing T7-polymerase (multiplicity of infection, MOI = 5), and cultured overnight in DMEM 10% FCS. Rifampicin was added to reduce cytopathicity. Effector cells (1×10^5) were mixed with target cells in a total volume of 250 μ l, allowed to fuse for 6 h and then lysed with 180 μ l of lysis buffer (20 mM Tris pH 7.5, 100 mM NH_4SO_4 , 0.1% BSA, 0.75% Triton X-100, and 0.001% sodium azide). Luciferase activity was detected by adding equal volumes (100 μ l) of luciferase detection reagent (Promega) and lysate, and the relative light units (r.l.u.) were determined by luminometry. When Raji B cells were used as target cells, 2.5×10^6 cells were transfected with 6 μ g of the indicated plasmid using Amaxa's Nucleofector electroporation device for transfection in buffer V or T on program A23. These suspension target cells were added onto the adherent PK13 effector cells prepared as described. Fusion was analysed as above.

Binding of soluble NiV-G to ephrinB2. 3 μ g ml⁻¹ of ephrinB1-Fc and ephrinB2-Fc diluted in ELISA buffer (2% BSA and 0.05% Tween-20 in TBS) were captured by biotinylated anti-human Fc (Caltag) pre-bound to NeutrAvidin-coated polystyrene plates (Pierce). Supernatant from sNiV-G-HA- or Δ 28sNiV-G-HA-transfected 293T cells was added to each well, and detected with an HRP-conjugated anti-HA antibody (Novus Biologicals) using TMB substrate (Pierce).

Infection assay. The VSV- Δ G-RFP virus is a recombinant VSV derived from a full-length complementary DNA clone of the VSV Indiana serotype in which the G-protein gene has been replaced with the RFP gene²⁰ (a gift from M. Whitt). Either VSV-G or NiV-F/G was provided *in trans*. NiV-F/G and VSV-G pseudotypes were purified via centrifugation through a sucrose cushion and used to infect 293T, Vero, CHO-pgsA745, rat cortical neurons and HMVECs (MOI = 1, as titred on 293T cells). RFP production at 24 h was analysed by fluorescent microscopy or FACS.

Neutralization sera. For NiV, New Zealand White rabbits were genetically immunized with a mixture of codon-optimized NiV-M (matrix), NiV-F and NiV-G expression plasmids (Aldevron) using an electroporation protocol that results in increased antibody titres²⁷. A 1:100 dilution of hyperimmune sera from the terminal bleed was used for neutralization studies. For VSV, a VSV-G-specific mouse monoclonal antibody (clone 8G5F II, a gift from J. Rose) was used. Pseudotyped viruses were pre-incubated with antibodies for 1 h before use for infection.

Received 18 March; accepted 17 May 2005.

Published online 6 July 2005.

1. Hsu, V. P. *et al.* Nipah virus encephalitis reemergence, Bangladesh. *Emerg. Infect. Dis.* 10, 2082–2087 (2004).

2. The International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDRDB). Person-to-person transmission of Nipah virus during outbreak in Faridpur District. *Health Sci. Bull.* **2**, 5–9 (2004).
3. Poliakov, A., Cotrina, M. & Wilkinson, D. G. Diverse roles of eph receptors and ephrins in the regulation of cell migration and tissue assembly. *Dev. Cell* **7**, 465–480 (2004).
4. Palmer, A. & Klein, R. Multiple roles of ephrins in morphogenesis, neuronal networking, and brain function. *Genes Dev.* **17**, 1429–1450 (2003).
5. Wong, K. T. *et al.* Nipah virus infection: pathology and pathogenesis of an emerging paramyxoviral zoonosis. *Am. J. Pathol.* **161**, 2153–2167 (2002).
6. Chua, K. B. *et al.* Nipah virus: a recently emergent deadly paramyxovirus. *Science* **288**, 1432–1435 (2000).
7. Harcourt, B. H. *et al.* Molecular characterization of Nipah virus, a newly emergent paramyxovirus. *Virology* **271**, 334–349 (2000).
8. Parashar, U. D. *et al.* Case-control study of risk factors for human infection with a new zoonotic paramyxovirus, Nipah virus, during a 1998–1999 outbreak of severe encephalitis in Malaysia. *J. Infect. Dis.* **181**, 1755–1759 (2000).
9. Lam, S. K. Nipah virus—a potential agent of bioterrorism? *Antiviral Res.* **57**, 113–119 (2003).
10. Field, H. *et al.* The natural history of Hendra and Nipah viruses. *Microbes Infect.* **3**, 307–314 (2001).
11. Wong, K. T. *et al.* A golden hamster model for human acute Nipah virus infection. *Am. J. Pathol.* **163**, 2127–2137 (2003).
12. Hooper, P., Zaki, S., Daniels, P. & Middleton, D. Comparative pathology of the diseases caused by Hendra and Nipah viruses. *Microbes Infect.* **3**, 315–322 (2001).
13. Bossart, K. N. *et al.* Membrane fusion tropism and heterotypic functional activities of the Nipah virus and Hendra virus envelope glycoproteins. *J. Virol.* **76**, 11186–11198 (2002).
14. Tamin, A. *et al.* Functional properties of the fusion and attachment glycoproteins of Nipah virus. *Virology* **296**, 190–200 (2002).
15. Guillaume, V. *et al.* Nipah virus: vaccination and passive protection studies in a hamster model. *J. Virol.* **78**, 834–840 (2004).
16. Gale, N. W. *et al.* Ephrin-B2 selectively marks arterial vessels and neovascularization sites in the adult, with expression in both endothelial and smooth-muscle cells. *Dev. Biol.* **230**, 151–160 (2001).
17. Shin, D. *et al.* Expression of ephrinB2 identifies a stable genetic difference between arterial and venous vascular smooth muscle as well as endothelial cells, and marks subsets of microvessels at sites of adult neovascularization. *Dev. Biol.* **230**, 139–150 (2001).
18. Rucker, J. *et al.* Cell-cell fusion assay to study role of chemokine receptors in human immunodeficiency virus type 1 entry. *Methods Enzymol.* **288**, 118–133 (1997).
19. Bossart, K. N. & Broder, C. C. Viral glycoprotein-mediated cell fusion assays using vaccinia virus vectors. *Methods Mol. Biol.* **269**, 309–332 (2004).
20. Takada, A. *et al.* A system for functional analysis of Ebola virus glycoprotein. *Proc. Natl Acad. Sci. USA* **94**, 14764–14769 (1997).
21. Esko, J. D., Stewart, T. E. & Taylor, W. H. Animal cell mutants defective in glycosaminoglycan biosynthesis. *Proc. Natl Acad. Sci. USA* **82**, 3197–3201 (1985).
22. Liu, J. & Thorp, S. C. Cell surface heparan sulfate and its roles in assisting viral infections. *Med. Res. Rev.* **22**, 1–25 (2002).
23. Kullander, K. & Klein, R. Mechanisms and functions of Eph and ephrin signalling. *Nature Rev. Mol. Cell Biol.* **3**, 475–486 (2002).
24. Noren, N. K., Lu, M., Freeman, A. L., Koolpe, M. & Pasquale, E. B. Interplay between EphB4 on tumour cells and vascular ephrin-B2 regulates tumour growth. *Proc. Natl Acad. Sci. USA* **101**, 5583–5588 (2004).
25. Estus, S. *et al.* Aggregated amyloid-beta protein induces cortical neuronal apoptosis and concomitant “apoptotic” pattern of gene induction. *J. Neurosci.* **17**, 7736–7745 (1997).
26. Shao, R. & Guo, X. Human microvascular endothelial cells immortalized with human telomerase catalytic protein: a model for the study of *in vitro* angiogenesis. *Biochem. Biophys. Res. Commun.* **321**, 788–794 (2004).
27. Tollefsen, S. *et al.* DNA injection in combination with electroporation: a novel method for vaccination of farmed ruminants. *Scand. J. Immunol.* **57**, 229–238 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Whitt for permission to use the VSV-pseudotype system, B. Reversade for discussions, and K. Adams for editorial comments. This work was supported by NIH grants to B.L., an NIH NRSA grant to O.A.N., an emerging infectious disease training grant to A.B.-C., and a biodefence research fellowship to E.L.L. S.T. was supported by an NSF-funded UCLA-IGERT bioinformatics traineeship. B.L. is a recipient of the Burroughs Wellcome Fund Career Development Award and is also a Charles E. Culpepper Medical Scholar supported by the Rockefeller Brothers Fund and by Goldman Philanthropic Partnerships. We also acknowledge support from the UCLA AIDS Institute and the flow cytometry core (UCLA CFAR).

Author Contributions E.L.L., H.C.A., A.B.-C. and R.N. contributed equally to this work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.L. (bleebhl@ucla.edu).

LETTERS

Regulation of *Mycobacterium tuberculosis* cell envelope composition and virulence by intramembrane proteolysis

Hideki Makinoshima¹ & Michael S. Glickman^{1,2}

Mycobacterium tuberculosis infection is a continuing global health crisis that kills 2 million people each year¹. Although the structurally diverse lipids of the *M. tuberculosis* cell envelope each have non-redundant roles in virulence or persistence^{2–7}, the molecular mechanisms regulating cell envelope composition in *M. tuberculosis* are undefined. In higher eukaryotes, membrane composition is controlled by site two protease (S2P)-mediated cleavage of sterol regulatory element binding proteins^{8,9}, membrane-bound transcription factors that control lipid biosynthesis. S2P is the founding member of a widely distributed family of membrane metalloproteases^{10,11} that cleave substrate proteins within transmembrane segments¹². Here we show that a previously uncharacterized *M. tuberculosis* S2P homologue (Rv2869c) regulates *M. tuberculosis* cell envelope composition, growth *in vivo* and persistence *in vivo*. These results establish that regulated intramembrane proteolysis is a conserved mechanism controlling membrane composition in prokaryotes and show that this proteolysis is a proximal regulator of cell envelope virulence determinants in *M. tuberculosis*.

Despite the well-established role of S2P in lipid metabolism in higher eukaryotes, prokaryotic S2P family members that have so far been characterized control sporulation in *Bacillus subtilis* (SpoIVFB)^{13,14}, the periplasmic stress response in *Escherichia coli* (YaeL)^{15,16}, and cell polarity¹⁷. To examine the physiological role of regulated intramembrane proteolysis (RIP) in *M. tuberculosis*, we searched the *M. tuberculosis* genome for S2P homologues with both the signature HEXxH zinc-chelation active site motif and the LDG carboxy-terminal motif present within predicted transmembrane domains. Through this approach we identified an S2P homologue (Rv2869c) in the *M. tuberculosis* chromosome that has not previously been characterized. Figure 1a shows the hydropathy plots of human S2P, YaeL and Rv2869c. Although the amino acid identities between the three proteins are low (16–22%), the conserved HEXxH and F/LDG motifs and transmembrane topology establish Rv2869c as an intramembrane cleaving protease (iCLIP)¹².

To characterize the function of Rv2869c in mycobacteria, we deleted this gene from the chromosomes of *M. bovis* BCG and *M. tuberculosis* Erdman by specialized transduction¹⁸. Figure 1b shows the genomic location of Rv2869c between *dxr* and *gcpE*, two genes in the non-mevalonate pathway of isoprenoid biosynthesis^{19,20}. Southern blotting (Fig. 1c) confirmed the successful replacement of Rv2869c with a hygromycin resistance cassette in both BCG and *M. tuberculosis*, showing that, in contrast to YaeL in *E. coli*, Rv2869c is not an essential gene for *M. tuberculosis* viability.

When grown on solid medium, the Rv2869c-null mutant had an altered colony morphology. In pathogenic mycobacteria, the colonial and microscopic morphology of cording has long been associated

with virulence and is dependent on multiple cell-envelope lipids^{2,5}. Both the BCG and *M. tuberculosis* Δ Rv2869c strains lacked cording, as measured by colonial morphology (Fig. 2a) and by microscopic examination of auramine–rhodamine-stained bacilli (Fig. 2b). Wild-type cording was restored to the Δ Rv2869c strain by a wild-type copy of Rv2869c, confirming that loss of Rv2869c function was responsible for the non-cording phenotype (Fig. 2). Alanine substitution mutations in the conserved HEXxH (H21A) and FDG (D340A) motifs, which are required for the proteolytic activity of RIP proteases^{10,13,16}, did not restore wild-type colony morphology (Fig. 2c). These results show that the proteolytic activity of Rv2869c regulates cell envelope composition in pathogenic mycobacteria. These results indicated that RIP might be a conserved mechanism of membrane composition control in prokaryotes.

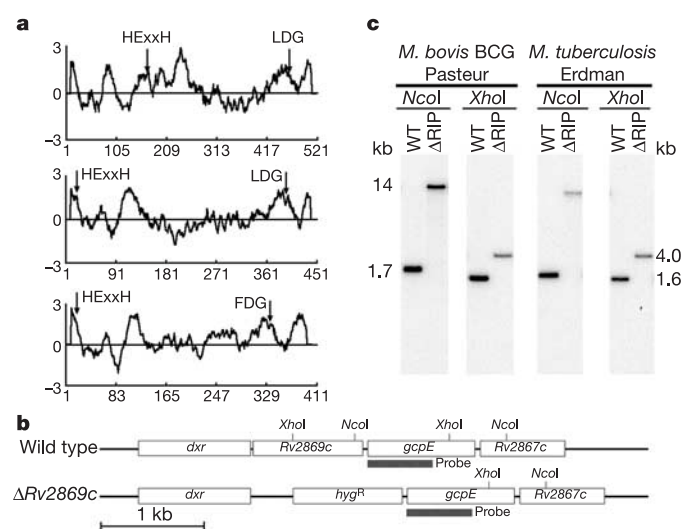


Figure 1 | Rv2869c is a non-essential intramembrane cleaving protease (iCLIP) of pathogenic mycobacteria. **a**, Hydropathy plots of human S2P (top), *E. coli* YaeL (middle) and *M. tuberculosis* Rv2869c (bottom), in which positive numbers on the y axis indicate hydrophobic residues and negative numbers indicate hydrophilic residues plotted against amino acid number on the x axis. Arrowheads indicate the conserved sequences (HExxH and F/LDG) within predicted transmembrane domains typical of intramembrane cleaving proteases. **b**, Map of the Rv2869c genomic region in the wild type and the Rv2869c mutant. Restriction sites and probe location are indicated. **c**, Southern blot of genomic DNA from indicated strains probed with the DNA fragment indicated in **b** showing allelic exchange at the Rv2869c locus in both *M. tuberculosis* and *M. bovis* BCG. WT, wild type.

¹Immunology Program, Sloan-Kettering Institute, and ²Division of Infectious Diseases, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA.

The sterol regulatory element binding protein (SREBP) pathway regulates multiple pathways of lipid biosynthesis, including cholesterol and fatty acids²¹. To examine directly whether *Rv2869c* regulates lipid composition in *M. tuberculosis*, we analysed the extractable and esterified mycolic acids of the cell envelope of wild-type cells, $\Delta Rv2869c$ cells and complemented mutant cells by quantitative thin-layer chromatography (TLC) of [¹⁴C]acetate-labelled cells in the presence or absence of detergent in the culture medium. In the esterified mycolic acids we observed similar quantities of the three major mycolic acids in all strains regardless of culture conditions (Fig. 3a, upper panel). In contrast, whereas wild-type cells maintained extractable mycolic acid synthesis after a shift to detergent-free medium (Fig. 3a, lower panel), $\Delta Rv2869c$ cells downregulated α -mycolate synthesis 4.6-fold, methoxymycolates 3.5-fold and ketomycolates 2.3-fold. In addition, $\Delta Rv2869c$ cells upregulated a slow-migrating lipid near the origin of the TLC plate by sixfold, which was absent from wild-type or complemented mutant extractable lipids (Fig. 3a, lower panel). These results indicated that *Rv2869c* regulates the composition of extractable mycolic acids in the cell envelope in response to changes in membrane fluidity but has no role in the covalently esterified mycolates of the cell wall. To examine whether *Rv2869c* also regulates other classes of cell envelope lipids, we examined the composition of phosphatidylinositol mannoside (PIM). Two-dimensional TLC of chloroform-methanol extracts revealed typical pattern of di- and hexamannosylated PIM (PIM2 and PIM6), with differing degrees of acylation^{22,23}. We found no difference between mutant and wild-type cells in PIM2 species (Supplementary Fig. S1, spots 1 and 2) but observed half the abundance of a PIM6 subspecies (spot 4) in mutant but not wild-type or complemented mutant cells.

To investigate whether *Rv2869c* controls cell envelope composition through transcriptional regulation, we compared the transcriptional

profiles of the wild-type and $\Delta Rv2869c$ strains in the presence or absence of detergent, using an *M. tuberculosis* whole-genome oligonucleotide microarray. We found that *Rv2869c* was both a positive and negative transcriptional regulator of multiple lipid biosynthetic and lipid-degrading genes. Consistent with the lipid analysis was our observation that *Rv2869c* mediated complex transcriptional regulation of mycolic acid biosynthetic genes in response to detergent, including *pks13*, *kasB*, *accD6*, *fabG1*, *fabG2* and the genes encoding fatty acid synthase and two lipid desaturases (Fig. 3b and Supplementary Information). Cluster analysis grouped the mycolic acid biosynthetic genes *kasB*, *pks13* and *accD6* together but indicated that other components of the pathway such as *fabG1/fabG2* were divergently regulated by *Rv2869c* (Fig. 3, Supplementary Fig. S2 and Supplementary Tables S1 and S2). Other lipid biosynthetic genes were positively regulated by *Rv2869c*, including multiple genes involved in PDIM synthesis (*mas*, *ppsA* and *rrrB*). The absence of *Rv2869c* resulted in strong overexpression of a lipase (*lipP*) and an epoxide hydrolase (*ephC*) and strong underexpression of *rpfC* (Supplementary Tables S1 and S2), three genes putatively involved in lipid or cell-wall degradation²⁴. Taken together, these data show that the lipid perturbations in the cell envelope of the *Rv2869c* mutant resulted from altered transcriptional control of diverse lipid anabolic and catabolic pathways.

An emerging model in *M. tuberculosis* pathogenesis is that the extractable lipids of the cell envelope act as direct effectors of pathogenesis either to modulate host immune responses or to alter intracellular trafficking^{2,3,7,25–29}. Some of these lipids are regulated by *Rv2869c*, indicating that RIP might be a proximal regulator of critical cell-envelope determinants of pathogenesis. To test this idea, we examined the *Rv2869c* mutant in the mouse model of aerosol

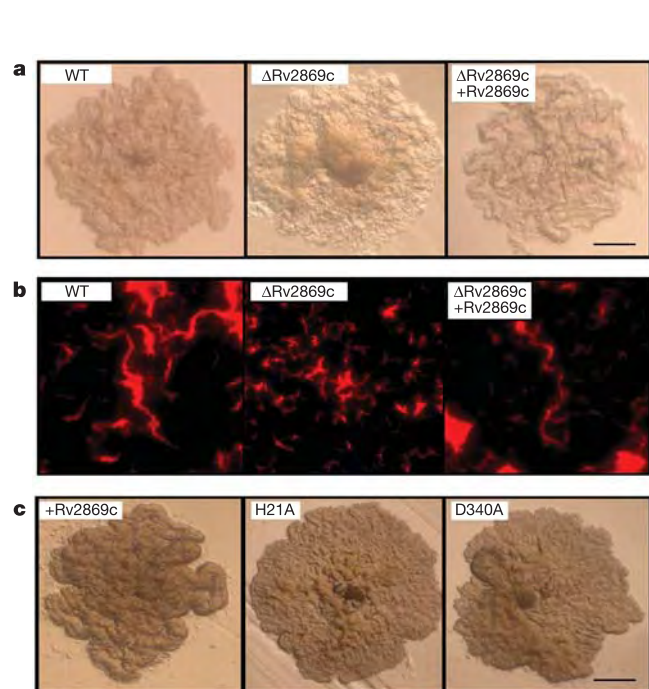


Figure 2 | *Rv2869c* and its proteolytic activity are required for mycobacterial cording. **a**, Single colonies of the indicated BCG strains at $\times 5$ magnification. Scale bar, 2 mm. WT, wild type. **b**, Auramine–rhodamine-stained smears of the indicated *M. tuberculosis* strains examined by fluorescence microscopy at $\times 400$ magnification. **c**, Conserved HExxH and FDG motifs of *Rv2869c* are required for complementation of the mutant cording phenotype. The *M. bovis* BCG *Rv2869c*-null mutant was complemented either with the wild-type *Rv2869c*, *Rv2869c*(H21A) or *Rv2869c*(D340A), and single colonies were assessed for the return of wild-type colony morphology.

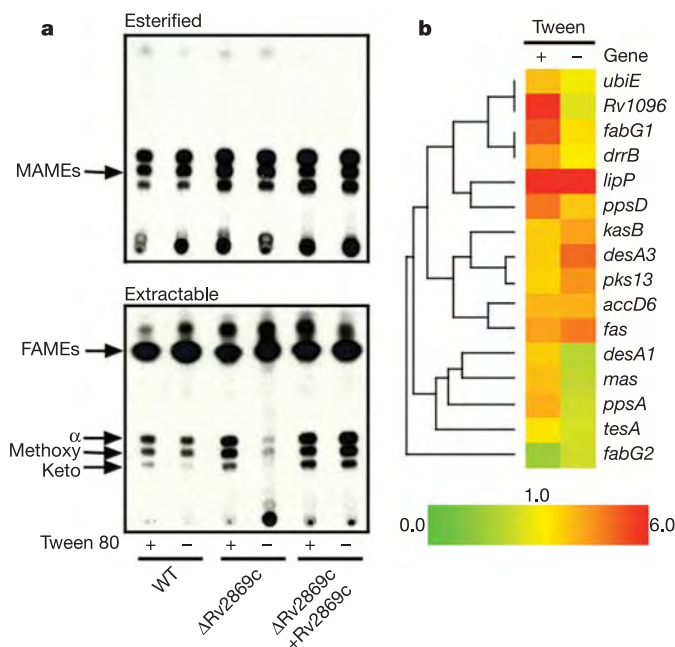


Figure 3 | *Rv2869c* transcriptionally regulates the extractable mycolic acid composition of the mycobacterial cell envelope. **a**, Radio-TLC of mycolic acid methyl esters (MAME) and fatty acid methyl esters (FAME) in the presence or absence of Tween 80. The top panel shows lipids covalently bound to the cell wall (esterified); the bottom panel shows extractable lipids. The three major mycolic acid subtypes (α , methoxy and ketomycolates) are labelled in the bottom panel. WT, wild type. **b**, *Rv2869c* controls cell envelope composition through positive and negative transcriptional control of lipid anabolic and catabolic genes. Lipid-related genes significantly regulated in the *Rv2869c* mutant as determined by microarray analysis in detergent (+ Tween) or short-term detergent withdrawal (– Tween) conditions. The colour bar depicts the ratio of mutant to wild type.

infection with *M. tuberculosis*. Despite identical inocula one day after infection, the *Rv2869c* mutant was impaired for bacterial growth during the acute phase of infection such that mutant bacterial titres in the lung after 3 weeks of infection were about 100-fold lower than in the wild type (Fig. 4a). In mice, wild-type *M. tuberculosis* persists at a constant titre in the lung for the life of the animal. We found that *Rv2869c* was also required for this persistence phase of the infection. $\Delta Rv2869c$ organisms were progressively eliminated from the lung such that by 22 weeks after infection, the number of bacilli in the lung infected with *Rv2869c* mutant was 10,000-fold lower than in the wild type. This *in vivo* phenotype was due to a loss of *Rv2869c* function and not to a polar effect on neighbouring genes, because the wild-type growth and persistence phenotypes were restored by a plasmid expressing *Rv2869c* (Fig. 4b). In the liver, $\Delta Rv2869c$ organisms were completely cleared by 22 weeks of infection (Supplementary Fig. S3). Gross and microscopic examination of infected lungs revealed a dramatic attenuation of granulomatous histopathology in the mutant-infected lungs (Fig. 4c, d).

Although S2P-mediated RIP of SREBPs is a well-described mechanism controlling eukaryotic membrane composition, S2P family members in prokaryotes that have so far been characterized regulate stress responses^{15,16,30}, sporulation¹³ and cell polarity¹⁷. The present study implicates RIP in controlling the lipid composition of the complex mycobacterial cell envelope. Consistent with the role of other RIP proteases in controlling transcription are the observed changes in cell envelope lipids that were associated with transcriptional dysregulation of lipid biosynthetic and degradative genes. Further mechanistic studies will determine whether *Rv2869c* directly cleaves membrane-bound transcriptional regulators or controls

membrane composition through cleavage of other membrane proteins that control membrane composition.

The recent sequencing of the *M. tuberculosis* genome has stimulated rapid progress in the identification of virulence determinants, many of which are involved in cell envelope biosynthesis. Although each of these individual lipid species serves distinct pathogenetic function, the cell envelope dysregulation in the *Rv2869c* mutant affected all phases of murine infection, including both *in vivo* replication and *in vivo* persistence. These results indicate that *Rv2869c* might control multiple cell envelope based virulence determinants. This study thus identifies regulated intramembrane proteolysis as an attractive target for *M. tuberculosis* drug development. Further characterization of mycobacterial RIP will yield important information about the regulation of the unique cell envelope of *M. tuberculosis* and the molecular mechanisms that allow *M. tuberculosis* to persist despite host immunity.

METHODS

Media and strains. *M. tuberculosis* strain Erdman was grown at 37 °C in 7H9 (broth) or 7H10 (agar) (Difco) medium supplemented with 10% oleic acid/albumin/dextrose/catalase (OADC), 0.5% glycerol, 0.05% Tween 80 (broth), and where appropriate hygromycin (Roche) at 50 $\mu\text{g ml}^{-1}$. BCG strains were grown similarly except that ADS supplement was substituted for OADC. Strain names are as follows: MGM307, BCG Pasteur *Rv2869c::hyg*; MGM309 = *M. tuberculosis* Erdman *Rv2869c::hyg*; MGM336 = MGM307 + pHMG121; MGM350 = MGM309 + pHMG121.

Construction of *Rv2869c*-null mutant by allelic change and complementation. Gene disruption was performed by specialized transduction of *Rv2869c::hygR* cassettes with temperature-sensitive mycobacteriophages as described previously^{2,18}. The *Rv2869c::hyg*-null allele included the first six nucleotides of *Rv2869c* (5') and the last 26 nucleotides of *Rv2869c* (3') flanking a hygromycin resistance cassette. After 4 weeks of selection on Middlebrook 7H10 agar medium containing 50 $\mu\text{g ml}^{-1}$ hygromycin, *hyg*^R transductants were genotyped at the *Rv2869c* locus by Southern blotting as described previously².

For complementation of the *Rv2869c* mutant, we reconstructed the *Rv2869c* operon with a truncated 55-amino-acid in-frame fusion of *Rv2870c* followed by the intact *Rv2869c* gene expressed from its native promoter 5' of *Rv2870c*. This complementation construct was used to transform MGM307 and MMG309 to kanamycin resistance.

Preparation and analysis of mycolates. Extractable lipids were prepared by the Folch method, dried under nitrogen and dissolved in diethyl ether. TLC was performed on Adsorbosil silica HPTLC plates (Alltech) and spots were revealed with 20% sulphuric acid in ethanol and charring. For PIM analysis, total lipids were developed with chloroform/methanol/water (60:30:6) as the first dimension and chloroform/acetic acid/methanol/water (40:25:3:6) as second dimension. For analysis of bound and extractable mycolic acids from *M. tuberculosis*, the indicated strains were labelled with [¹⁴C]acetic acid and mycolic acids were prepared as described previously² from chloroform/methanol extracts (extractable) and cell pellets (cell-wall bound). TLC development conditions were three developments with hexanes/ethyl acetate (95:5). Radio-TLCs were imaged by phosphorimaging, and quantification of individual spots was performed on a Fuji FLA-5000 and Image Gauge software (Version 4.1, Fujifilm).

Animal Infections. Before infection, exponentially replicating bacteria ($D_{600} = 0.3$) were washed in PBS containing 0.05% Tween 80, and sonicated to disperse clumps. For aerosol infection, mice were exposed to 8×10^7 colony-forming units (c.f.u.) of the appropriate strain in a Middlebrook Inhalation Exposure System (Glas-Col). This dose of bacteria delivers 100 c.f.u. per animal. Bacterial burden was determined by plating serial dilutions of lung, spleen or liver homogenates on 7H10 agar plates. Plates were incubated at 37 °C in 5% CO₂ for 3–6 weeks before colonies were counted. For histological analysis, organs were fixed in 10% normal buffered formalin and embedded in paraffin; 6- μm sections were stained with haematoxylin and eosin.

Microarray analysis. An *M. tuberculosis* whole-genome microarray was produced with the Operon *Tuberculosis* whole-genome oligo set, version 1.0 (Operon). Details of probe preparation, hybridization and microarray data have been deposited in the Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo/) under accession number GSE2561. This complete array was printed on Corning UltraGaps coated slides with a Microgrid TAS arrayer (Genomic Solutions) in duplicate such that each slide contained two complete copies of the array. Hybridizations were performed on biological triplicates (containing Tween, six total spots per gene) or duplicates (in the absence of Tween, four total spots per gene) and included one dye-reversal experiment for

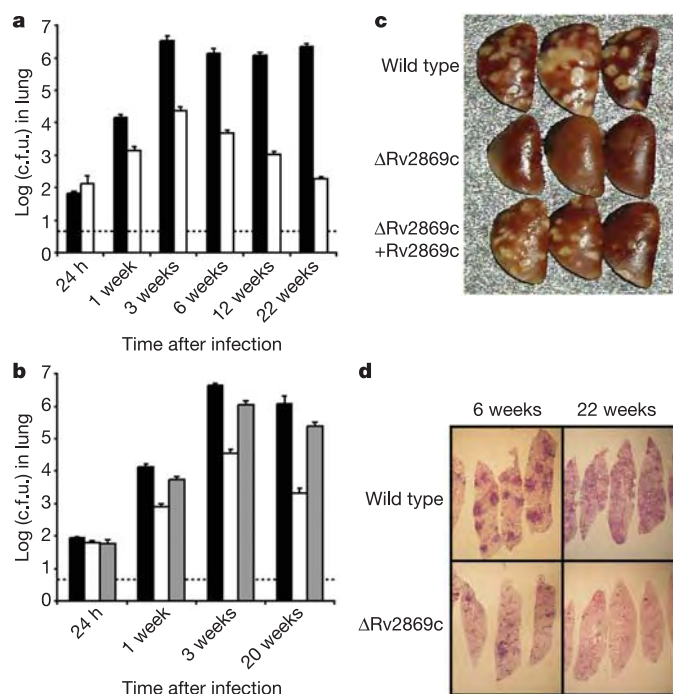


Figure 4 | *Rv2869c* is required for both *M. tuberculosis* replication and persistence in mice. **a, b**, Lung bacterial loads plotted (logarithmic scale) from mice infected by aerosol with wild-type *M. tuberculosis* (black bars), the $\Delta Rv2869c$ strain (open bars) and the complemented mutant (grey bars) in two separate experiments. The horizontal line indicates the limit of detection of the assay; error bars are s.d. **c**, Gross pathology of infected lungs at 20 weeks from the wild type (top), the *Rv2869c* mutant (middle) or the complemented mutant (bottom). **d**, Haematoxylin/eosin-stained lung tissue from C57BL/6 mice infected with the wild type (upper panels) and the *Rv2869c* mutant (lower panels) at 6 weeks (left panels) and 22 weeks (right panels) after infection.

each condition. Spot quantification was performed in GenePix 4.0 and data analysis was performed in GeneSpring 7.0 with intensity-dependent Lowess normalization. Overexpression of *Rv2868c* and *Rv2867c* were excluded from the list of regulated genes because of the possibility that this observation was a polar effect of the hygromycin marker.

Active-site mutagenesis of *Rv2869c*. Site-directed mutagenesis was performed on *Rv2869c* by the overlapping extension mutagenesis method on pHMG121. Substitution H21A or D340A was generated with mutagenic primers HMG123 (5'-CCCACATGTGACACATTCGGCCAGGGCCACCGAAATCAGG-3') and HMG124 (5'-CCTGATTCGGTGGCCCTGGCCGAATGTGGTCACATGTGGG-3') or HMG127 (5'-GCGACGGCAATATGGCCGCCGGCGAACGGCA GCAACGGCAG-3') and HMG128 (5'-CTGCCGTGTCTGCCGTTCGCCGGC GGCCATATTGCCGTCCG-3'), respectively. The fragment 5' of the H21A mutation was a 176-base-pair (bp) product of the primers HMG109 and HMG123, and the 3' fragment was a 451-bp product of primers HMG124 and HMG112. In a third PCR the overlapping fragments were mixed and amplified into a 588-bp DNA by using the primers HMG109 and HMG110. The 5' fragment for D340A mutation a 553-bp product with the primers HMG113 and HMG127, and the 3' fragment was amplified as a 214-bp product with the primers HMG113 and HMG110. In a third PCR the overlapping fragments were mixed and amplified into a 736-bp DNA by using the primers HMG113 and HMG110. Fragments containing the point mutations were sequenced completely and subcloned in pHMG121 at the *Xba*I and *Eco*RI sites for H21A and at the *Eco*RI and *Hind*III sites for D340A.

Received 21 January; accepted 3 May 2005.

- WHO. *Global Tuberculosis Control: Surveillance, Planning, Financing* (World Health Organization, Geneva, 2004).
- Glickman, M. S., Cox, J. S. & Jacobs, W. R. Jr. A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*. *Mol. Cell* **5**, 717–727 (2000).
- Reed, M. B. et al. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* **431**, 84–87 (2004).
- Cox, J. S., Chen, B., McNeil, M. & Jacobs, W. R. Jr. Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. *Nature* **402**, 79–83 (1999).
- Gao, L. Y. et al. Requirement for *kasB* in *Mycobacterium* mycolic acid biosynthesis, cell wall impermeability and intracellular survival: implications for therapy. *Mol. Microbiol.* **49**, 1547–1563 (2003).
- Dubnau, E. et al. Oxygenated mycolic acids are necessary for virulence of *Mycobacterium tuberculosis* in mice. *Mol. Microbiol.* **36**, 630–637 (2000).
- Rao, V., Fujiwara, N., Porcelli, S. A. & Glickman, M. S. *Mycobacterium tuberculosis* controls host innate immune activation through cyclopropane modification of a glycolipid effector molecule. *J. Exp. Med.* **201**, 535–543 (2005).
- Brown, M. S., Ye, J., Rawson, R. B. & Goldstein, J. L. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. *Cell* **100**, 391–398 (2000).
- Sakai, J. et al. Sterol-regulated release of SREBP-2 from cell membranes requires two sequential cleavages, one within a transmembrane segment. *Cell* **85**, 1037–1046 (1996).
- Rawson, R. B. et al. Complementation cloning of S2P, a gene encoding a putative metalloprotease required for intramembrane cleavage of SREBPs. *Mol. Cell* **1**, 47–57 (1997).
- Duncan, E. A., Dave, U. P., Sakai, J., Goldstein, J. L. & Brown, M. S. Second-site cleavage in sterol regulatory element-binding protein occurs at transmembrane junction as determined by cysteine panning. *J. Biol. Chem.* **273**, 17801–17809 (1998).
- Weihofen, A. & Martoglio, B. Intramembrane-cleaving proteases: controlled liberation of proteins and bioactive peptides. *Trends Cell Biol.* **13**, 71–78 (2003).
- Rudner, D. Z., Fawcett, P. & Losick, R. A family of membrane-embedded metalloproteases involved in regulated proteolysis of membrane-associated transcription factors. *Proc. Natl Acad. Sci. USA* **96**, 14765–14770 (1999).
- Cutting, S., Roels, S. & Losick, R. Sporulation operon *spo*VF and the characterization of mutations that uncouple mother-cell from forespore gene expression in *Bacillus subtilis*. *J. Mol. Biol.* **221**, 1237–1256 (1991).
- Alba, B. M., Leeds, J. A., Onufryk, C., Lu, C. Z. & Gross, C. A. DegS and YaeL participate sequentially in the cleavage of RseA to activate the σ^E -dependent extracytoplasmic stress response. *Genes Dev.* **16**, 2156–2168 (2002).
- Kanehara, K., Ito, K. & Akiyama, Y. YaeL (EcfE) activates the σ^E pathway of stress response through a site-2 cleavage of anti- σ^E , RseA. *Genes Dev.* **16**, 2147–2155 (2002).
- Chen, J. C., Viollier, P. H. & Shapiro, L. A membrane metalloprotease participates in the sequential degradation of a *Caulobacter* polarity determinant. *Mol. Microbiol.* **55**, 1085–1103 (2005).
- Bardarov, S. et al. Specialized transduction: an efficient method for generating marked and unmarked targeted gene disruptions in *Mycobacterium tuberculosis*, *M. bovis* BCG and *M. smegmatis*. *Microbiol.* **148**, 3007–3017 (2002).
- Kuzuyama, T., Takahashi, S., Takagi, M. & Seto, H. Characterization of 1-deoxy-D-xylulose 5-phosphate reductoisomerase, an enzyme involved in isopentenyl diphosphate biosynthesis, and identification of its catalytic amino acid residues. *J. Biol. Chem.* **275**, 19928–19932 (2000).
- Altincicek, B. et al. GcpE is involved in the 2-C-methyl-D-erythritol 4-phosphate pathway of isoprenoid biosynthesis in *Escherichia coli*. *J. Bacteriol.* **183**, 2411–2416 (2001).
- Horton, J. D. et al. Combined analysis of oligonucleotide microarray data from transgenic and knockout mice identifies direct SREBP target genes. *Proc. Natl Acad. Sci. USA* **100**, 12027–12032 (2003).
- Gillerson, M., Quesniaux, V. F. & Puzo, G. Acylation state of the phosphatidylinositol hexamannosides from *Mycobacterium bovis* Bacillus Calmette Guérin and *Mycobacterium tuberculosis* H37Rv and its implication in Toll-like receptor response. *J. Biol. Chem.* **278**, 29880–29889 (2003).
- Kremer, L. et al. Characterization of a putative alpha-mannosyltransferase involved in phosphatidylinositol trimannoside biosynthesis in *Mycobacterium tuberculosis*. *Biochem. J.* **363**, 437–447 (2002).
- Cohen-Gonsaud, M. et al. The structure of a resuscitation-promoting factor domain from *Mycobacterium tuberculosis* shows homology to lysozymes. *Nature Struct. Mol. Biol.* **12**, 270–273 (2005).
- Glickman, M. S. & Jacobs, W. R. Jr. Microbial pathogenesis of *Mycobacterium tuberculosis*: dawn of a discipline. *Cell* **104**, 477–485 (2001).
- Schaible, U. E. et al. Apoptosis facilitates antigen presentation to T lymphocytes through MHC-I and CD1 in tuberculosis. *Nature Med.* **9**, 1039–1046 (2003).
- Rhoades, E. et al. Identification and macrophage-activating activity of glycolipids released from intracellular *Mycobacterium bovis* BCG. *Mol. Microbiol.* **48**, 875–888 (2003).
- Vergne, I. et al. *Mycobacterium tuberculosis* phagosome maturation arrest: mycobacterial phosphatidylinositol analog phosphatidylinositol mannoside stimulates early endosomal fusion. *Mol. Biol. Cell* **15**, 751–760 (2004).
- Hingley-Wilson, S. M., Sambandamurthy, V. K. & Jacobs, W. R. Jr. Survival perspectives from the world's most successful pathogen, *Mycobacterium tuberculosis*. *Nature Immunol.* **4**, 949–955 (2003).
- Schobel, S., Zellmeier, S., Schumann, W. & Wiegert, T. The *Bacillus subtilis* sigmaW anti-sigma factor RsiW is degraded by intramembrane proteolysis through YluC. *Mol. Microbiol.* **52**, 1091–1105 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Bongiorno and F. Gao for technical support; V. Rao, N. Serbina, P. Wong and N. Stephanou for discussions; and A. Viale and the Genomics Core Lab of the Memorial Sloan-Kettering Cancer Center for assistance with microarray experiments. M.S.G. is supported by an NIH grant, the Ellison Medical Foundation, and the Speakers Fund for Biomedical Research awarded by the City of New York.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S.G. (glickmam@mskcc.org).

LETTERS

Trans-SNARE pairing can precede a hemifusion intermediate in intracellular membrane fusion

Christoph Reese¹, Felix Heise² & Andreas Mayer¹

The question concerning whether all membranes fuse according to the same mechanism has yet to be answered satisfactorily. During fusion of model membranes or viruses, membranes dock, the outer membrane leaflets mix (termed hemifusion), and finally the fusion pore opens and the contents mix^{1,2}. Viral fusion proteins consist of a membrane-disturbing 'fusion peptide' and a helical bundle that pin the membranes together²⁻⁴. Although SNARE (soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor) complexes form helical bundles with similar topology, it is unknown whether SNARE-dependent fusion events on intracellular membranes proceed through a hemifusion state. Here we identify the first hemifusion state for SNARE-dependent fusion of native membranes, and place it into a sequence of molecular events: formation of helical bundles by SNAREs precedes hemifusion; further progression to pore opening requires additional peptides. Thus, SNARE-dependent fusion may proceed along the same pathway as viral fusion: both use a docking mechanism via helical bundles^{5,6} and additional peptides to destabilize the membrane and efficiently induce lipid mixing⁷⁻⁹. Our results suggest that a common lipidic intermediate³ may underlie all fusion reactions of lipid bilayers.

We dissected membrane docking, lipid flow and fusion pore opening in a SNARE-dependent fusion reaction using the cell-free fusion of yeast vacuoles as a model¹⁰. To monitor lipid transition between fusing vacuoles we incorporated self-quenching concentrations of rhodamine-labelled phosphatidylethanolamine (Rh-PE) into isolated vacuolar membranes (see Supplementary Information). Rh-PE-labelled vacuoles were mixed with a fivefold excess of unlabelled vacuoles. Fusion of labelled with unlabelled vacuoles dilutes Rh-PE over a larger surface, resulting in de-quenching and an increase in relative fluorescence.

Upon start of a fusion reaction, we observed a nonlinear ATP-dependent increase in Rh-PE fluorescence (Fig. 1a). This increase depended on dilution by fusion with unlabelled vacuoles because it was absent if unlabelled acceptor vacuoles were omitted (not shown). In the presence of acceptor vacuoles but in the absence of ATP—a condition that inactivates the authentic fusion pathway to more than 95%¹¹—fluorescence showed a slow linear increase. This may represent spontaneous fusion-independent dye transfer or be due to unincorporated dye micelles sticking to re-isolated vacuoles, which we consider as more likely. Such material could provide a source for fusion-independent dye incorporation into unlabelled vacuoles during the incubation. This background signal could be clearly distinguished from authentic fusion by testing fusion-incompetent mutants and established fusion inhibitors.

First, we blocked early steps of vacuole fusion, priming and docking. Priming activates SNAREs via the ATPase Sec18/NSF and its cofactor Sec17/ α -SNAP¹⁰. We cultivated strains carrying temperature-sensitive alleles of Sec18/NSF (*sec18-1*) or

Sec17/ α -SNAP (*sec17-1*)¹² at permissive temperature. During vacuole isolation, they were briefly shifted to restrictive temperature to rapidly inactivate Sec18 and Sec17. In fusion reactions neither *sec18-1* nor *sec17-1* vacuoles showed an ATP-dependent increase in

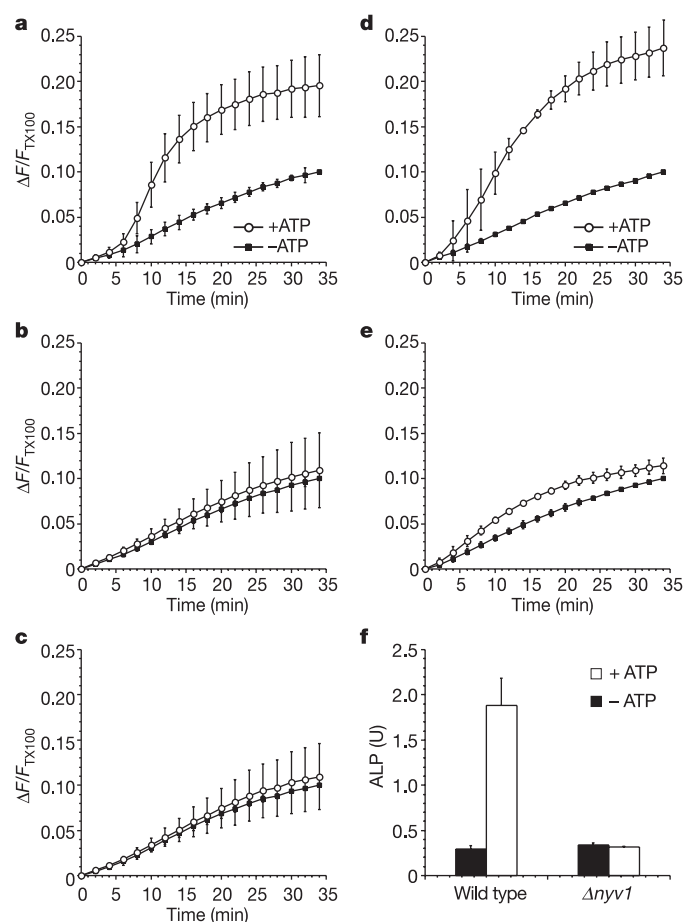


Figure 1 | Lipid mixing between vacuoles from priming and docking mutants. **a–c**, Strains carrying wild-type *SEC18* and *SEC17* (**a**), or the temperature-sensitive alleles *sec17-1* (**b**) or *sec18-1* (**c**), were grown at 25 °C and briefly shifted to 37 °C during vacuole isolation. Lipid mixing was assayed as described in Methods in fusion reactions run in the presence or absence of ATP. **d–f**, Vacuoles were isolated from wild-type *NYV1* (**d**) or isogenic $\Delta nyv1$ (**e**) cells. Lipid mixing (**d**, **e**) and alkaline phosphatase (ALP)-based content-mixing (**f**) were assayed from fusion samples as in **a–c**. Data represent the means of four independent experiments with standard deviation (s.d.).

¹Département de Biochimie, Université de Lausanne, Chemin des Boveresses 155, CH-1066 Epalinges, Switzerland. ²Biochemie-Zentrum der Universität Heidelberg (BZH), Im Neuenheimer Feld 328, D-69120 Heidelberg, Germany.

relative fluorescence that was comparable to the isogenic wild type (Fig. 1b, c). Furthermore, vacuoles isolated from strains lacking the vacuolar v-SNARE Nyv1, which is required for docking, did not show a significant ATP-dependent increase in relative fluorescence (Fig. 1d, e). We simultaneously assayed content mixing in these samples using the conventional vacuolar fusion assay¹⁰. This assay is based on maturation of pro-alkaline phosphatase in the lumen of one fusion partner by transfer of a maturation enzyme from the lumen of the second vacuole. The lack of Rh-PE de-quench correlated to a lack of alkaline phosphatase signal in the content mixing assay (Fig. 1f).

Antibodies to fusion-relevant proteins, inhibitory proteins acting on fusion factors, and low-molecular-mass compounds can inhibit vacuole fusion¹⁰. We tested antibodies to Sec18 and Sec17 that block priming. Both suppressed lipid- and content mixing (Fig. 2). Control IgG molecules isolated from pre-immune serum had no influence. Gdi1 (guanine nucleotide dissociation inhibitor), which extracts the Rab-GTPase Ypt7 from vacuolar membranes and blocks vacuole tethering¹⁰, thoroughly inhibited both lipid- and content mixing. The strong effects of inactivating vacuolar SNAREs or the vacuolar Rab-GTPase indicate that the Rh-PE assay detects the authentic SNARE- and Rab-dependent fusion pathway.

Proteoliposome reconstitution experiments suggested that trans-SNARE pairing induces lipid mixing¹³. In that case, trans-SNARE pairing should be inseparable from lipid mixing. Otherwise, treatments that block transition from membrane docking to content mixing might permit trans-SNARE pairing but prevent lipid flow. To distinguish these two possibilities we incubated fusion reactions of wild-type vacuoles in the presence of two inhibitors of the post-docking phase, the Ca^{2+} chelator BAPTA^{10,14,15} or affinity purified antibodies to calmodulin. These reagents inhibited both lipid- and content mixing (Fig. 2). Alternatively, we ran fusion reactions without inhibitors but using vacuoles lacking Vph1 ($\Delta vph1$). Vph1 is a subunit of the V_0 sector of the vacuolar V-type ATPase, which is required for the post-docking phase of a vacuole fusion reaction^{16,17}. $\Delta vph1$ vacuoles gave neither a lipid- nor a content-mixing signal (Fig. 3), even if we added an excess of the peripheral vacuolar t-SNARE subunit Vam7, which can bypass some early requirements of vacuole fusion¹⁷. Also, deletion of another late-acting factor of vacuole fusion, Vac8 (ref. 18), compromised lipid- and content mixing to a comparable extent (not shown). BAPTA and lysophosphatidylcholine, a lipid that modifies bilayer conformation, as well as deletion of Vph1 or Vac8, permit trans-SNARE pairing^{10,16–19} but block lipid mixing. Therefore, trans-SNARE pairing may not

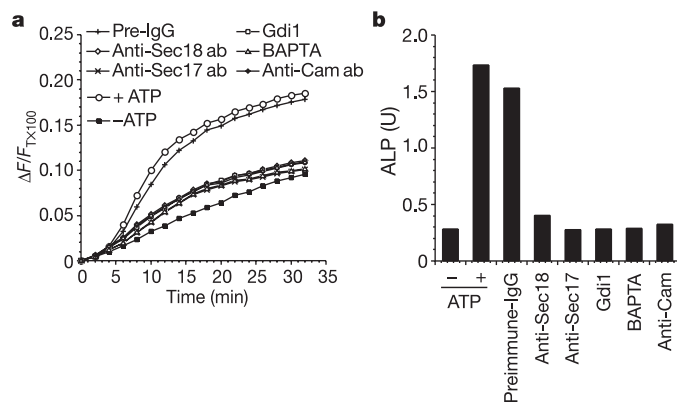


Figure 2 | Inhibitor sensitivity of lipid mixing. Wild-type vacuoles were prepared, stained with Rh-PE and used in fusion reactions containing Gdi1 (4 μM), BAPTA (5 mM), antibodies to Sec18 (0.3 μM), Sec17 (0.3 μM), calmodulin (1.5 μM), or non-immune antibodies (3 μM), or buffer only. One sample was incubated without ATP. Fusion was followed by fluorescence de-quenching of Rh-PE (a) or the ALP based content-mixing assay (b).

suffice to induce lipid flow. This transition requires the BAPTA-sensitive step and Vph1.

A critical question for understanding the fusion pathway is whether lipids merge independently of fusion-pore opening. We explored this issue by titrating inhibitors into fusion samples and simultaneously assaying lipid- and content mixing (Fig. 4a–c). Because both assays operate under identical reaction conditions, equal concentrations of inhibitors must yield equal absolute fusion activities. Titrating the response of both assays thus allows the calibration of the two signals against each other. Signals in both assays were severely reduced by antibodies to Sec18/NSF (inhibiting priming), the t-SNARE Vam3 (inhibiting docking), or the protein phosphatase 1 Glc7 (inhibiting post-docking²⁰). Because equivalent inhibitor concentrations reduced the Rh-PE signals slightly less than the ALP signals, the lipid-mixing assay is more sensitive to small fusion activities. All agents used obviously affected steps before lipid mixing and could not dissociate lipid- from content mixing.

A qualitatively different response was obtained for the post-docking inhibitor GTP- γS ^{11,16,21}. GTP- γS had little influence on lipid mixing up to 4 mM but it suppressed content mixing with a half-maximal inhibitory concentration (IC_{50}) of 1 mM (Fig. 4d; see also Supplementary Fig. 2). Lipid mixing in the presence of GTP- γS was sensitive to inactivation of the vacuolar fusion machinery by BAPTA, anti-Vam3 antibody, anti-Sec18 antibody, Gdi1, or by deletion of Nyv1 or Vph1 (Fig. 4f), confirming that it resulted from authentic fusion. Thus, GTP- γS arrested the reaction at an intermediate stage, permitting lipid flow between the fusion partners but no content mixing. This arrest was reversible because GTP- γS -blocked vacuoles completed content mixing after re-isolation and removal of GTP- γS (Fig. 4e).

The content-mixing assay requires transfer of a soluble maturase (Pep4; >45 kDa). We sought to test whether lipid flow might occur through a smaller protein-impermeable pore by loading vacuoles with the small, soluble fluorophore calcein²². Calcein (molecular mass 622 Da) has the same molecular backbone as the headgroup of

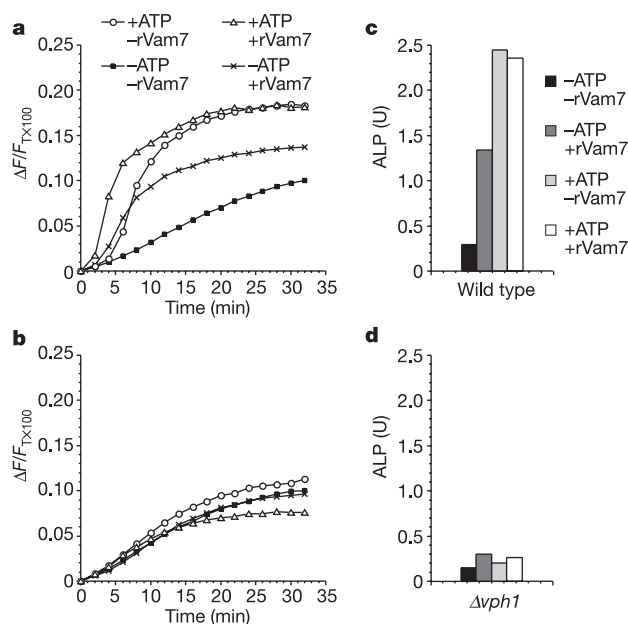


Figure 3 | Lipid mixing in the vacuolar V_0 knockout $\Delta vph1$. a, b, Vacuoles were prepared from wild type (a) or isogenic $\Delta vph1$ (b) strains. The membranes were stained with Rh-PE and used in fusion reactions with the indicated combinations of ATP and recombinantly expressed and purified t-SNARE subunit Vam7 (0.4 μM). Lipid mixing was measured via Rh-PE de-quenching. c, d, Aliquots of the same samples as in a and b were incubated in parallel and used to determine content mixing via ALP (c, wild type; d, $\Delta vph1$).

Rh-PE that we used as a lipid reporter (sulphorhodamine B; molecular mass 559 Da; Supplementary Fig. 4). Under conditions where calcein transfer does not occur, Rh-PE de-quench thus indicates lipid flow between the cytoplasmic membrane leaflets in the absence of full fusion and content mixing.

We loaded vacuoles with calcein and re-isolated them by flotation. They were mixed with a fivefold excess of unlabelled vacuoles and incubated under fusion conditions. In the case of complete fusion both diameter and frequency of the calcein-labelled vacuoles should increase. In the case of small pores, which pass calcein but do not expand, the diameter of the calcein vacuoles should remain constant but their frequency should increase. If lipid flow on at least the inner leaflet were blocked neither diameter nor frequency should increase. After a fusion reaction, both diameter and frequency of calcein-labelled vacuoles had increased in an ATP-dependent fashion, which was sensitive to BAPTA and Gd1 (Fig. 5; see also Supplementary

Fig. 3). GTP- γ S prevented the increase in size and frequency as efficiently as BAPTA and Gd1 (Fig. 5; see also Supplementary Fig. 3), demonstrating that GTP- γ S prevents even small fusion pores. Because GTP- γ S permitted unrestricted Rh-PE de-quench (Fig. 4d) we conclude that this inhibitor arrested vacuole fusion at a hemifusion state.

We operationally define hemifusion as lipid flux in the absence of content mixing. Vacuolar hemifusion follows the BAPTA-sensitive step but precedes the GTP- γ S-sensitive step (Supplementary Fig. 5). Vph1 and, by inference, vacuolar V_0 were required to induce lipid mixing. V_0 undergoes a transformation between the BAPTA- and GTP- γ S-sensitive steps that can be diagnosed by formation of V_0 - V_0 complexes between two fusing organelles. It preferentially affects V_0 complexes associated with the t-SNARE Vam3 (refs 16,21). Several observations connect V_0 to the induction of lipid flow. First, V_0 transformation is sensitive to lysophosphatidylcholine¹⁹, a lipid that modifies bilayer conformation²³. Second, lipid transition kinetically maps to the same stage as V_0 transformation. Third, inactivation of V_0 permits trans-SNARE pairing but blocks lipid transition. Finally, mutation of the *Drosophila* Vph1 homologue blocks fusion of synaptic vesicles between docking and pore opening²⁴. On the basis of these observations we propose that trans-SNARE pairing forces the bilayers into close contact and triggers downstream events, such as Ca^{2+} efflux^{14,15} and V_0 activation^{16,17,21}. Efficient lipid transition and pore opening require additional fusogenic influences, which, based on the kinetic analyses, might depend on V_0 .

Similarly, type I viral fusion proteins contain separable activities for pinning the membranes together and for destabilizing the bilayer². Type I fusion proteins insert a fusion peptide into the target membrane and fold back onto themselves to form a helical bundle holding the membranes in proximity⁵. The nature of the fusion peptide is crucial to the triggering of fusion. Single-amino-acid

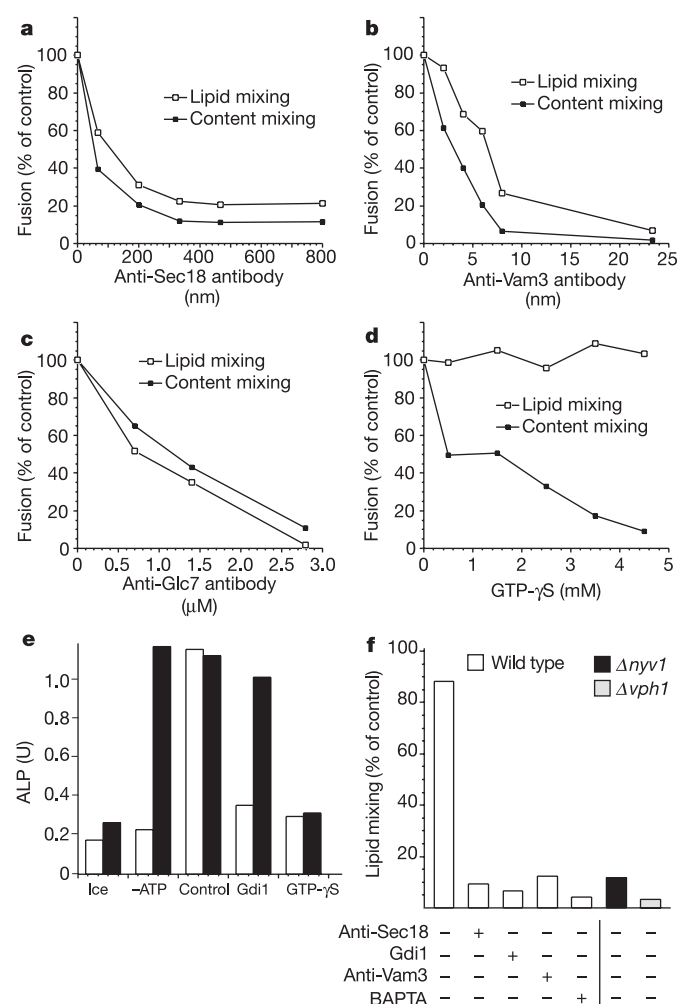


Figure 4 | Sensitivity of lipid and content mixing to inhibitors. **a–d**, Lipid and content mixing were assayed in fusion reactions containing titrations of antibodies to Sec18 (**a**), antibodies to Vam3 (**b**), antibodies to Glc7 (**c**), or GTP- γ S (**d**). Values obtained without inhibitors served as the 100% reference. **e**, Reversibility of the GTP- γ S block. Fusion reactions with 4.5 mM GTP- γ S, containing Rh-PE-labelled vacuoles in the same ratio as in **a–d**, were run for 40 min. After re-isolation to remove GTP- γ S, vacuoles were incubated for another 70 min at 27 °C. Content mixing was assayed (filled bars). In a parallel set of samples, fusion was run (70 min) without GTP- γ S and re-isolation (open bars). **f**, Fusion reactions containing wild type, $\Delta nyv1$ or $\Delta vph1$ vacuoles were run in the presence of 6 mM GTP- γ S and in addition either BAPTA (5 mM), anti-Vam3, anti-Sec18, Gd1, or buffer as above. Lipid mixing was assayed after 30 min. The 100% reference contained no inhibitors.

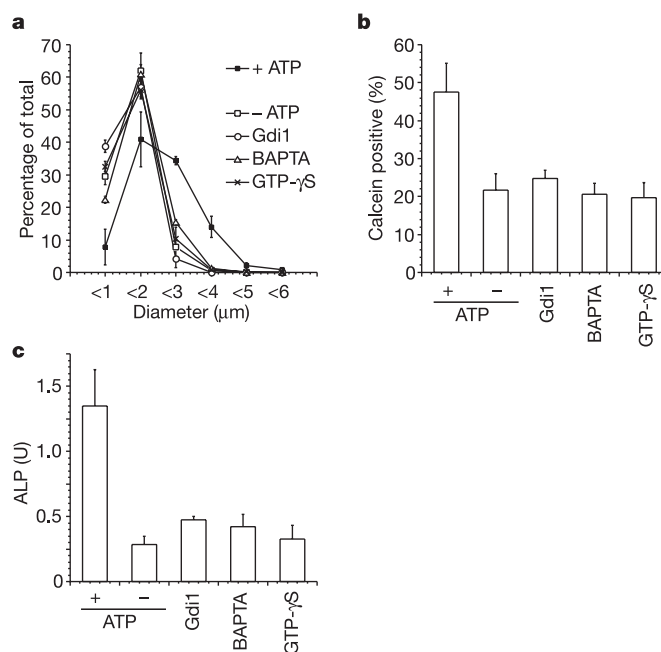


Figure 5 | Morphometric analysis of content mixing. Vacuoles were loaded with calcein-AM, re-isolated and mixed with a fivefold excess of unlabelled vacuoles. Fusion reactions were run in the presence of GTP- γ S (4 mM), BAPTA (5 mM), or Gd1 (4 μ M). After 40 min, vacuoles were counterstained with BODIPY- C_{12} (2 μ M) and analysed by confocal microscopy. **a**, **b**, Diameters (**a**) and frequencies (**b**) of calcein-labelled vacuoles were measured for >1,000 vacuoles per experiment and condition. Three independent experiments were averaged. Error bars indicate the s.d. **c**, An aliquot of each sample was used to assay content mixing via ALP activity.

substitutions in the fusion peptide block fusion without compromising helical bundle formation^{3,4,7,8,25}. On the other hand, soluble fusion peptides alone can fuse liposomes⁴; that is, their 'fusogenicity' is independent of membrane anchoring. Results from vacuole fusion²⁶, viral fusion^{2–5,7,8,25} and proteoliposome fusion⁹ are consistent with the idea that separable activities provide a mechanochemical device for membrane apposition and a fusogenic influence. These separable activities are joined into a single viral fusion protein but may be provided by separate polypeptides in SNARE-dependent fusion. In agreement with this, SNAREs can be proteolytically removed from sea urchin secretory vesicles^{27,28}. This does not compromise their ability to fuse in a Ca^{2+} -dependent fashion as long as the membranes are docked by other means. In viral fusion, transition from hemifusion to pore opening is considered as the most energy-demanding step. The same may apply to vacuole fusion. Even reactions released from a GTP- γ S block, resuming fusion after lipid transition, do not progress faster than a complete reaction¹⁹. Thus, in vacuole fusion the rate-limiting step also appears to follow hemifusion, providing another mechanistic analogy.

In this study we have identified the first hemifusion state in a physiological SNARE-dependent fusion reaction. Hemifusion is not compatible with simple versions of a gap junction type of pore model of membrane fusion^{29,30}, although the model can be modified to accommodate the existence of hemifusion. In contrast, a hemifusion intermediate is the hallmark of stalk models of fusion. Stalk models describe the fusion of model membranes and viruses well^{1,3}. They predict that the fusion site is formed by lipids and that the stalk is formed by locally restricted fusion of the outer membrane leaflets. The fact that SNARE-dependent fusion also comprises a hemifusion state suggests that all fusion reactions might proceed along a similar pathway. How hemifusion is induced in a controlled fashion and what drives pore expansion remain key questions for the future.

METHODS

All methods not listed here can be found in Supplementary Information.

Rh-PE labelling of vacuoles. Rhodamine labelled 1,2-dihexadecanoyl-sn-glycero-3-phosphoethanolamine (Rh-PE, Molecular Probes) does not readily dissolve in DMSO. We added 1.25 ml of analytical grade DMSO (Sigma-Aldrich, D8418) per 5 mg Rh-PE (3 mM solution) and incubated the preparation (1 h, 37 °C) under occasional vortexing. Once Rh-PE had completely dissolved, 80- μ l aliquots were prepared and stored in siliconized 1.5-ml reaction tubes (Biozym) at –20 °C. Rh-PE aliquots were thawed and shaken at 37 °C and 1,400 r.p.m. for 20 min. Rh-PE aliquots were centrifuged (15 min at room temperature, 13,100 g) to pellet aggregated Rh-PE. The supernatant was used immediately for labelling. A total of 560 μ g of freshly prepared vacuoles were equilibrated to 32 °C in 800 μ l PS buffer (10 mM PIPES/KOH pH 6.8, 200 mM sorbitol) in siliconized 1.5-ml reaction tubes for approximately 40 s at 500 r.p.m. Sixty microlitres of the Rh-PE solution were withdrawn, carefully avoiding the pelleted Rh-PE, and added to the equilibrated vacuoles in a drop-wise fashion under gentle vortexing. The tube was incubated in a water bath at 27 °C for 30 s. A total of 500 μ l of PS buffer with 15% (w/v) Ficoll (pre-warmed to 27 °C) was added, the suspension was gently mixed and transferred to a siliconized 2-ml reaction tube (pre-cooled on ice). For pipetting, 1-ml Gilson tips were cut open in order to minimize shear forces on the vacuolar membranes. Vacuoles were overlaid with 200 μ l of PS buffer containing 4% (w/v) Ficoll (27 °C) and 500 μ l PS buffer containing 0% (w/v) Ficoll (27 °C). The gradient was centrifuged (5 min, 3 °C, 11,700 g, swingout rotor) with slow acceleration and deceleration. Stained vacuoles were recovered from the 0%/4%-Ficoll interface.

Lipid-mixing assay. Sixfold reactions with a volume of 189 μ l and a final vacuole concentration of 0.2 mg ml^{–1} were set up: 120 μ l vacuole mastermix (0.3 mg ml^{–1}) in PS was supplemented with 1.3 mg ml^{–1} cytosol, 0.3 mM MnCl₂ and, finally, 110 mM KCl. Inhibitors were premixed in 60 μ l PS with 110 mM KCl and 0.3 mM MnCl₂. A total of 60 μ l inhibitor mix were added to 120 μ l vacuole mastermix, supplemented with 9.5 μ l of $\times 20$ ATP-regenerating system and gently vortexed. One-hundred microlitres of reaction mix were pipetted into non-coated black 96-well plates (no. 237105, NUNC) pre-cooled to 0 °C. Air bubbles were avoided by not completely ejecting the suspension from the tips. The microtitre plate was pre-treated immediately before use with 5% (w/v) skim-milk powder in water (1 h). The plate was washed, dried and cooled

on ice. Fluorescence change was measured with a SpectraMax GeminiXS fluorescent plate reader (Molecular Devices) at 27 °C, 538 nm excitation and 585 nm emission. Measurements were taken every 2 min for a total time of 32 min, yielding fluorescence values at the onset (F_0) and during the reaction (F_t). After completion of the reaction, 100 μ l of PS with 1% (w/v) Triton-X100 was added. Fluorescence was followed for 10 min, taking measurements every 30 s. The 20 measurements, which showed a small decrease, were averaged to yield fluorescence after infinite dilution ($F_{T\times 100}$). The relative fluorescence change $\Delta F_t/F_{T\times 100} = (F_t - F_0)/F_{T\times 100}$ was calculated for every time point t . $F_{T\times 100}$ was invariant over time; that is, $F_{T\times 100}$ values were comparable when Triton X-100 was added before or after the fusion reaction. Therefore, $F_{T\times 100}$ taken at the end of the fusion reaction was used as a reference for all time points. Addition of inhibitors to intact Rh-PE-labelled vacuoles did not change the fluorescence values. The $t = 32$ min value $\Delta F_t/F_{T\times 100}$ for the –ATP samples was set to 0.1 to facilitate comparison of the curves.

Labelling the vacuolar lumen with calcein-AM. Calcein can be taken up into the vacuolar lumen in its membrane-permeable acetoxymethyl ester form (calcein-AM). Vacuolar hydrolases cleave the ester bonds of calcein-AM, yielding its bright green fluorescent form, which is membrane impermeable. Hence the fusion strains rich in esterase (DKY6281) were labelled. To this end, vacuoles were prepared with the following slight modifications: cells were spheroplasted for 25 min in a final volume of 12 ml spheroplasting buffer containing 1 ml of lyticase and 200 μ l calcein-AM (1 mM in DMSO), yielding a final concentration of 17 μ M calcein-AM. After cell wall digestion and re-suspension of the cells in 2.5 ml PS buffer containing 15% (w/v) Ficoll 400, another 100 μ l of 1 mM calcein-AM in DMSO was added, yielding a final concentration of 40 μ M of calcein-AM. Vacuole isolation was continued as outlined in Supplementary Information.

Received 21 January; accepted 9 May 2005.

Published online 29 May 2005.

- Zimmerberg, J. & Chernomordik, L. V. Membrane fusion. *Adv. Drug Deliv. Rev.* **38**, 197–205 (1999).
- Cohen, F. S. & Melikyan, G. B. The energetics of membrane fusion from binding, through hemifusion, pore formation, and pore enlargement. *J. Membr. Biol.* **199**, 1–14 (2004).
- Chernomordik, L. V. & Kozlov, M. M. Protein-lipid interplay in fusion and fission of biological membranes. *Annu. Rev. Biochem.* **72**, 175–207 (2003).
- Tamm, L. K., Han, X., Li, Y. & Lai, A. L. Structure and function of membrane fusion peptides. *Biopolymers* **66**, 249–260 (2002).
- Eckert, D. M. & Kim, P. S. Mechanisms of viral membrane fusion and its inhibition. *Annu. Rev. Biochem.* **70**, 777–810 (2001).
- Sutton, R. B., Fasshauer, D., Jahn, R. & Brunger, A. T. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 Å resolution. *Nature* **395**, 347–353 (1998).
- Han, X., Bushweller, J. H., Cafiso, D. S. & Tamm, L. K. Membrane structure and fusion-triggering conformational change of the fusion domain from influenza hemagglutinin. *Nature Struct. Biol.* **8**, 715–720 (2001).
- Qiao, H., Armstrong, R. T., Melikyan, G. B., Cohen, F. S. & White, J. M. A specific point mutant at position 1 of the influenza hemagglutinin fusion peptide displays a hemifusion phenotype. *Mol. Biol. Cell* **10**, 2759–2769 (1999).
- Tucker, W. C., Weber, T. & Chapman, E. R. Reconstitution of Ca^{2+} -regulated membrane fusion by synaptotagmin and SNAREs. *Science* **304**, 435–438 (2004).
- Wickner, W. Yeast vacuoles and membrane fusion pathways. *EMBO J.* **21**, 1241–1247 (2002).
- Mayer, A., Wickner, W. & Haas, A. Sec18p (NSF)-driven release of Sec17p (α -SNAP) can precede docking and fusion of yeast vacuoles. *Cell* **85**, 83–94 (1996).
- Kaiser, C. A. & Schekman, R. Distinct sets of SEC genes govern transport vesicle formation and fusion early in the secretory pathway. *Cell* **61**, 723–733 (1990).
- Weber, T. et al. SNAREpins: minimal machinery for membrane fusion. *Cell* **92**, 759–772 (1998).
- Peters, C. & Mayer, A. Ca^{2+} /calmodulin signals the completion of docking and triggers a late step of vacuole fusion. *Nature* **396**, 575–580 (1998).
- Merz, A. J. & Wickner, W. T. Trans-SNARE interactions elicit Ca^{2+} efflux from the yeast vacuole lumen. *J. Cell Biol.* **164**, 195–206 (2004).
- Bayer, M. J., Reese, C., Buhler, S., Peters, C. & Mayer, A. Vacuole membrane fusion: V0 functions after trans-SNARE pairing and is coupled to the Ca^{2+} -releasing channel. *J. Cell Biol.* **162**, 211–222 (2003).
- Thorngren, N., Collins, K. M., Fratti, R. A., Wickner, W. & Merz, A. J. A soluble SNARE drives rapid docking, bypassing ATP and Sec17/18p for vacuole fusion. *EMBO J.* **23**, 2765–2776 (2004).
- Wang, Y. X., Kauffman, E. J., Duex, J. E. & Weisman, L. S. Fusion of docked membranes requires the armadillo repeat protein Vac8p. *J. Biol. Chem.* **276**, 35133–35140 (2001).

19. Reese, C. & Mayer, A. The rate limiting step of vacuolar membrane fusion follows trans-SNARE pairing and lipid-mixing. *J. Cell Biol.* (submitted).
20. Peters, C. *et al.* Control of the terminal step of intracellular membrane fusion by protein phosphatase 1. *Science* **285**, 1084–1087 (1999).
21. Peters, C. *et al.* Trans-complex formation by proteolipid channels in the terminal phase of membrane fusion. *Nature* **409**, 581–588 (2001).
22. Wang, L., Seeley, E. S., Wickner, W. & Merz, A. J. Vacuole fusion at a ring of vertex docking sites leaves membrane fragments within the organelle. *Cell* **108**, 357–369 (2002).
23. Vogel, S. S., Leikina, E. A. & Chernomordik, L. V. Lysophosphatidylcholine reversibly arrests exocytosis and viral fusion at a stage between triggering and membrane merger. *J. Biol. Chem.* **268**, 25764–25768 (1993).
24. Hiesinger, P. R. *et al.* The v-ATPase VO subunit a1 is required for a late step in synaptic vesicle exocytosis in *Drosophila*. *Cell* **121**, 607–620 (2005).
25. Epan, R. M. Fusion peptides and the mechanism of viral fusion. *Biochim. Biophys. Acta* **1614**, 116–121 (2003).
26. Ungermann, C., Sato, K. & Wickner, W. Defining the functions of trans-SNARE pairs. *Nature* **396**, 543–548 (1998).
27. Coorsen, J. R. *et al.* Regulated secretion: SNARE density, vesicle fusion and calcium dependence. *J. Cell Sci.* **116**, 2087–2097 (2003).
28. Szule, J. A. *et al.* Calcium-triggered membrane fusion proceeds independently of specific presynaptic proteins. *J. Biol. Chem.* **278**, 24251–24254 (2003).
29. Lindau, M. & Almers, W. Structure and function of fusion pores in exocytosis and ectoplasmic membrane fusion. *Curr. Opin. Cell Biol.* **7**, 509–517 (1995).
30. Zimmerberg, J. How can proteolipids be central players in membrane fusion? *Trends Cell Biol.* **11**, 233–235 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Reinhardt, A. Schmidt and V. Comte for assistance, and N. Garin and M. Allegrini for help with confocal microscopy. This work was supported by grants from DFG, FNS, HFSP and Boehringer Ingelheim.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.M. (Andreas.Mayer@unil.ch).

Structural basis of family-wide Rab GTPase recognition by rabenosyn-5

Sudharshan Eathiraj¹, Xiaojing Pan¹, Christopher Ritacco^{1†} & David G. Lambright¹

Rab GTPases regulate all stages of membrane trafficking, including vesicle budding, cargo sorting, transport, tethering and fusion^{1,2}. In the inactive (GDP-bound) conformation, accessory factors facilitate the targeting of Rab GTPases to intracellular compartments^{3–8}. After nucleotide exchange to the active (GTP-bound) conformation, Rab GTPases interact with functionally diverse effectors including lipid kinases, motor proteins and tethering complexes. How effectors distinguish between homologous Rab GTPases represents an unresolved problem with respect to the specificity of vesicular trafficking. Using a structural proteomic approach, we have determined the specificity and structural basis underlying the interaction of the multivalent effector rabenosyn-5 with the Rab family. The results demonstrate that even the structurally similar effector domains in rabenosyn-5 can achieve highly selective recognition of distinct subsets of Rab GTPases exclusively through interactions with the switch and interswitch regions. The observed specificity is determined at a family-wide level by structural diversity in the active conformation, which governs the spatial disposition of critical conserved recognition determinants, and by a small number of both positive and negative sequence determinants that allow further discrimination between Rab GTPases with similar switch conformations.

Many Rab GTPases, including Rab4, -5, -7, -9, -11, -14, -21 and -22, have overlapping localizations within the endosomal system^{2,9–15}. The sequential endocytic and recycling functions of Rab5 and Rab4 are further linked by multivalent effectors, such as rabenosyn-5, rabaptin-5 and Rabip4, which interact with both Rab proteins through separate domains^{16–18}. Multiple Rab partners have also been identified for other effectors^{13,19}. Although effectors engage conserved and non-conserved regions of Rab GTPases^{20–22}, what determines specificity at the Rab family level remains unknown.

To understand better how structural variability in the switch

regions contributes to Rab–effector recognition, crystal structures were determined for 11 Rab GTPases bound to GppNHp and/or GDP (Supplementary Fig. 1 and Supplementary Tables 1–3). Combined with earlier studies (Supplementary Table 3), at least one structure is available for the active conformation of 14 Rab GTPases, spanning six subfamilies^{23,24} and representing one-third of the Rab family, excluding isoforms. With the exception of Rab21, both the switch I and switch II regions adopt a unique active conformation independent of crystal packing (Supplementary Fig. 2). The inactive switch conformations, however, are either disordered or dictated by crystal contacts. In Rab21, the switch II region remains poorly ordered even in the active conformation.

To compare the active structures, superpositions were generated for all pair-wise combinations (Supplementary Fig. 3). Whereas the most similar elements coincide with conserved motifs required for nucleotide and Mg²⁺ binding, structural variability is a hallmark of the switch II region, the interswitch region and the $\alpha 3/\beta 5$ loop. Although the interswitch region and $\alpha 3/\beta 5$ loop are typically mobile, the structural diversity in the switch II region is primarily the consequence of non-conservative substitutions²⁵. Consistent with evolutionary pressure on functional sites, non-phylogenetic relationships occur at the local structural level. For example, the switch regions of Rab4 (subfamily II) adopt an active conformation similar to Rab5 and Rab22 (subfamily V). Conversely, the active switch conformations are dissimilar for Rab7 and Rab9 (subfamily VII) as well as Rab4 and Rab11 (subfamily II). These local relationships contribute to Rab–effector recognition and are reminiscent of the non-phylogenetic functional specificity reported for the Ras superfamily²⁶.

Rabenosyn-5 (Rbsn) contains distinct Rab4 and Rab5 binding sites within residues 264–500 and 627–784, respectively¹⁷. As shown in Fig. 1 and Supplementary Fig. 4, the Rab5 binding domain maps to

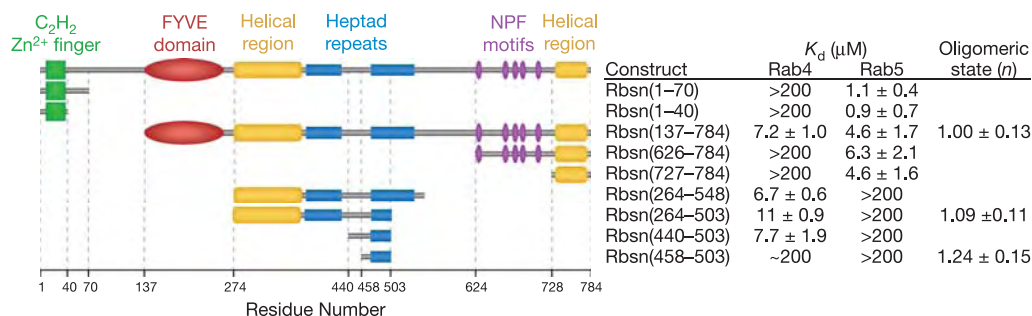


Figure 1 | Mapping of the rabenosyn-5 Rab GTPase binding domains. Mean K_d values and standard deviations were obtained from 2–4 surface plasmon resonance experiments using purified Rab GTPases and rabenosyn-5 constructs. The oligomeric state ($n = 1$ for monomer; $n = 2$ for dimer) was

determined from sedimentation equilibrium experiments over the concentration range from 5 to 50 μ M. Values represent the mean and standard deviation for three measurements.

¹Program in Molecular Medicine and Department of Biochemistry & Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA. [†]Present address: Molecular Biophysics & Biochemistry Department, Yale University, New Haven, Connecticut 06520, USA.

a predicted helical region at the carboxy terminus (Rbsn(728–784)). A BLAST search detected significant homology with residues 458–503 in the Rab4 binding region, although not with other proteins. Rbsn(458–503) overlaps the second of two heptad repeats, yet this region and Rbsn(728–784) are monomeric. Although Rbsn(458–503) binds Rab4, the affinity is 10–100-fold weaker than that of a longer construct, Rbsn(440–503), which represents the minimal Rab4 binding domain. The structure of Rbsn(458–503) reveals a helical hairpin similar to an anti-parallel coiled coil (Supplementary Fig. 4). In the complexes with Rab GTPases discussed below, an equivalent structure is observed for the homologous core. These observations raise two important questions: (1) how specific are the rabenosyn-5 Rab binding domains; and (2) how do they distinguish structurally similar Rab GTPases?

The specificities of Rbsn(440–503) and Rbsn(728–784) were profiled against 33 purified Rab GTPases (Fig. 2; see also Supplementary Fig. 5). Rbsn(440–503) binds with similar affinity to Rab4 and Rab14 and weakly to Rab2, whereas Rbsn(728–784) exhibits the highest affinity for Rab5, threefold lower affinity for Rab22 and Rab24, and binds weakly to Rab14. Other interactions are too weak for detection ($K_d > 200$ mM). Thus, Rbsn(440–503) and Rbsn(728–784) selectively recognize distinct subsets of Rab GTPases, despite similar core structures. Although further studies are required to determine the *in vivo* significance, it is noteworthy that both Rab14 and Rab22 extensively co-localize with Rab5 and have been implicated in endosomal trafficking^{12,13,27,28}. Given that the cellular functions of rabenosyn-5 and other effectors may be mediated by interactions with subsets of Rab GTPases, it will be of interest to consider the specificity for the entire Rab family when

evaluating the molecular basis by which effectors regulate trafficking events.

Crystallization screens for the high-affinity combinations yielded crystals for the Rab4–Rbsn(440–503) and Rab22–Rbsn(728–784) complexes. The structures reveal similar modes of interaction, with the helical hairpins engaging equivalent residues in the switch and interswitch regions (Fig. 3; see also and Supplementary Fig. 6). Surface areas of 2,136 Å² and 1,312 Å² are buried at the respective interfaces. The significantly larger contact area in the Rab4–Rbsn(440–503) complex evidently compensates for the weak binding of Rab4 to the helical hairpin core.

The binding epitope in Rbsn(728–784) is divided into two hydrophobic pockets by a polar ridge bisecting the long axis. One pocket buries invariant residues (Phe 42 and Trp 75) from the switch I and interswitch regions of Rab22. The second pocket, which is capped by residues 733–PIEEEEL-738 at the amino terminus of α 1, docks variable residues in switch II (Leu 70 and Met 73) and partially buries a cluster of conserved residues at the switch interface (Ile 38, Arg 66 and Phe 67). The side chains of residues lining the polar ridge mediate hydrogen-bonding interactions with the backbone of Ala 41 in switch I and the side-chain hydroxyl of Tyr 74 in switch II. Additional polar interactions at the periphery involve conserved (Lys 55) and variable (Met 73) residues in the interswitch and switch regions.

In Rbsn(440–503), the N terminus of α 1 is foreshortened by Pro 458, which packs against Val 73 in the switch II region of Rab4 and, in conjunction with an ionic interaction between Glu 454 and Arg 69, directs the N-terminal region into a flexible loop that connects with the β 1 strand encoded by residues 441–EGWLP-445.

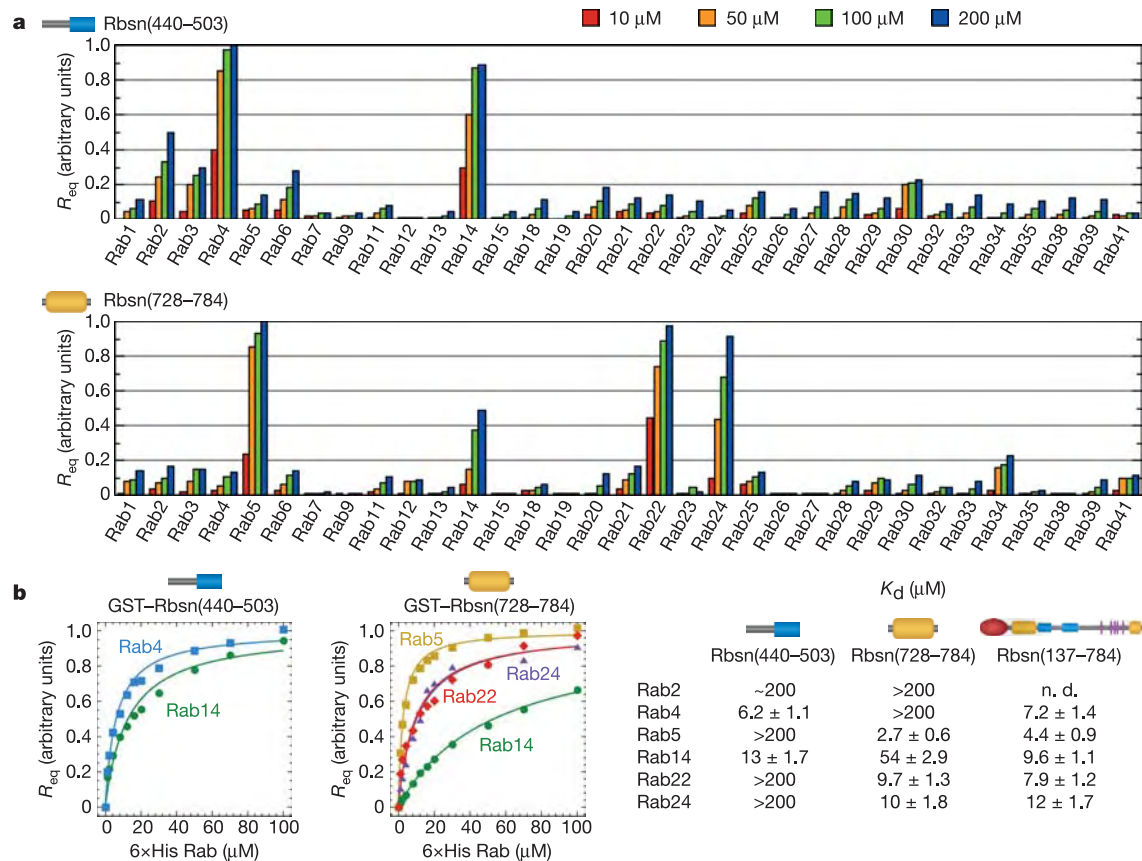


Figure 2 | Quantitative family-wide analysis of Rab GTPase-effector specificity. **a**, Initial screen for the interaction of 6 × His (or GST) fusions of Rab GTPases with GST (or 6 × His) fusions of Rbsn(440–503) and Rbsn(728–784). For each potential interaction, the equilibrium surface plasmon resonance signal (R_{eq}) was measured at four concentrations of the 6 × His Rab GTPase or 6 × His Rbsn construct. **b**, Concentration dependence of the equilibrium surface plasmon resonance signal (R_{eq}) for the binding of 6 × His Rab GTPases to GST fusions of Rbsn(440–503) and Rbsn(728–784). Mean K_d values and standard deviations for two to four independent experiments are tabulated on the right.

Residues 441–EGWLP–445 engage the $\beta 2$ strand of Rab4 via two main-chain hydrogen bonds, allowing Trp 443 and Pro 445 to pack against the invariant Phe 45 in switch I while Leu 444 occupies a hydrophobic pocket on the $\alpha 1$ helix of Rab4. This pocket exists as a consequence of a glycine residue following Phe 45. In most Rab GTPases, including Rab5 and Rab22, this glycine is replaced by a bulky side chain. Finally, several substitutions contribute to differences in the interface with the helical hairpin.

To identify compositional determinants of the observed specificity, we analysed the effects of mutations within or proximal to the binding interfaces (Fig. 4). In Rab5, broadly conserved residues were substituted with alanine, whereas residues selectively conserved in Rab5, Rab22 and Rab24 were replaced with the consensus residue for other Rab GTPases. Although most substitutions have little effect or otherwise enhance the affinity for Rbsn(728–784), a tenfold decrease in affinity is observed for the A57D and M89S substitutions. The corresponding residues in Rab22 (Ser 41 in switch I and Met 73 in switch II) are located within or proximal to the interface with Rbsn(728–784). Met 73 is buried in a hydrophobic pocket flanked by variable residues in the helical hairpin. Interestingly, the M89A substitution is without effect, suggesting that the M89S defect reflects sequestration of a polar residue in a non-polar environment. Conversely, the effects of the Ala 57 substitution correlate with side-chain volume rather than polarity, consistent with a packing defect.

The ability of Rbsn(728–784) to recognize Rab5 and Rab22 yet

discriminate against Rab21 derives, in part, from a distinctive substitution in Rab21 whereby a glutamine replaces the otherwise invariant switch I glycine. The corresponding G55Q substitution in Rab5 decreases affinity by two orders of magnitude. As the C α of Gly 55 packs against residues in switch II, a glutamine substitution should alter the structure and probably the stability of the active conformation. Indeed, the structure of the G55Q mutant reveals a main-chain perturbation in which Gln 55 bulges outward. Superposition with Rab22 suggests that van der Waals overlap contributes to the observed binding defect. Notably, the reverse Q53G substitution in Rab21 does not confer the ability to interact with Rbsn(728–784), nor is it sufficient to allow Rab21 to adopt an active conformation similar to Rab5 or Rab22 (Supplementary Fig. 2).

Order of magnitude decreases in binding affinity are also observed for the I734E, L738P and V764T mutations that replace residues in Rbsn(728–784) with the corresponding residues in Rbsn(440–503). Ile 734 and Leu 738 form one side of the non-polar pocket for residues Phe 67, Met 73 and Leu 70 in the switch II region of Rab22. Both polarity and packing are adversely affected by the I734E substitution, whereas the L738P substitution disrupts the main-chain conformation and side-chain packing. As Val 764 packs against Phe 42, Leu 57 and Trp 59 in the switch I and interswitch regions of Rab22, desolvation is a likely factor underlying the V764T defect.

Mutations in the N-terminal region of Rbsn(440–503) have only

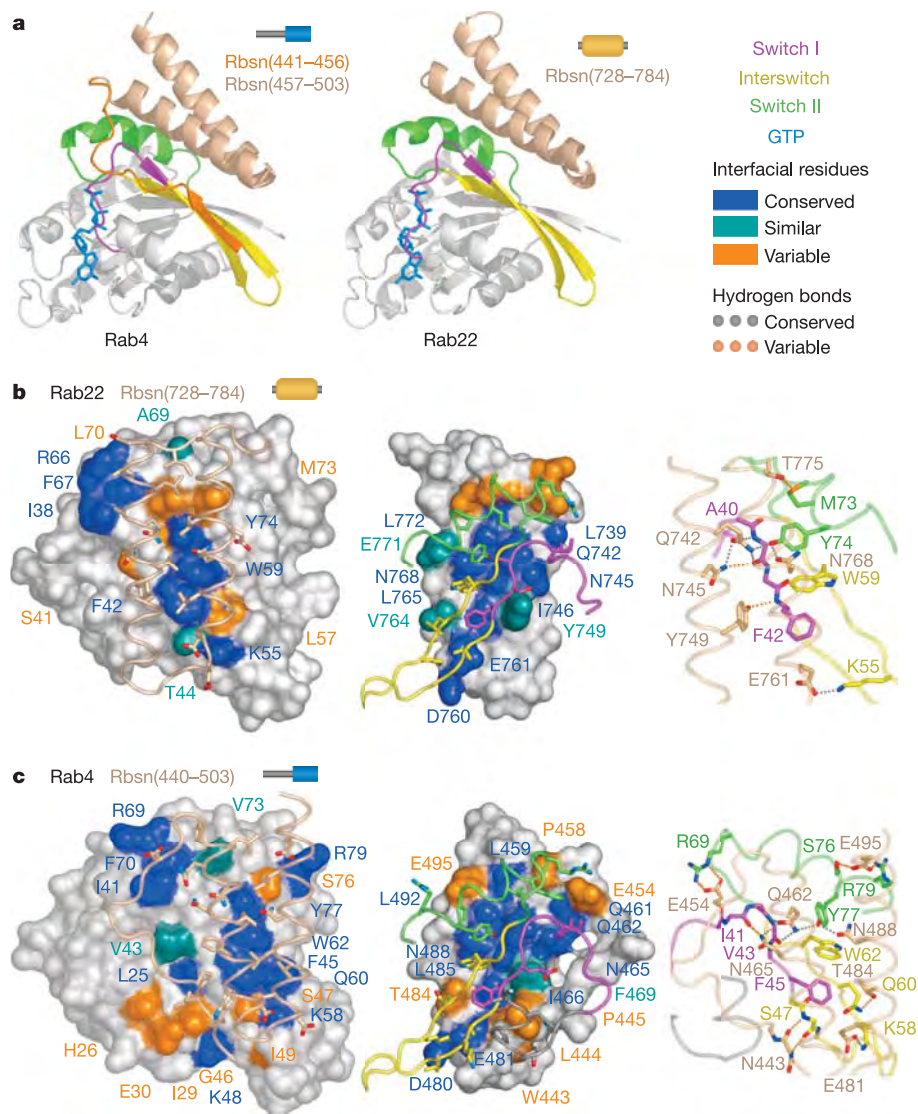


Figure 3 | Structural basis of Rab recognition by rabenosyn-5. **a**, Ribbon rendering of GTP-bound Rab4(Q67L) and Rab22(Q64L) in complex with the minimal Rab binding domains of rabenosyn-5. **b**, Conservation and variability in the Rab22–Rbsn(728–784) interface. Spheres covered by a semitransparent surface represent Rab22 (left panel) or Rbsn(728–784) (middle panel). Hydrogen-bonding interactions are depicted in the right panel. **c**, Conservation and variability in the Rab4–Rbsn(440–503) interface. Spheres covered by a semitransparent surface represent Rab4 (left panel) or Rbsn(440–503) (middle panel). Hydrogen-bonding interactions are depicted in the right panel.

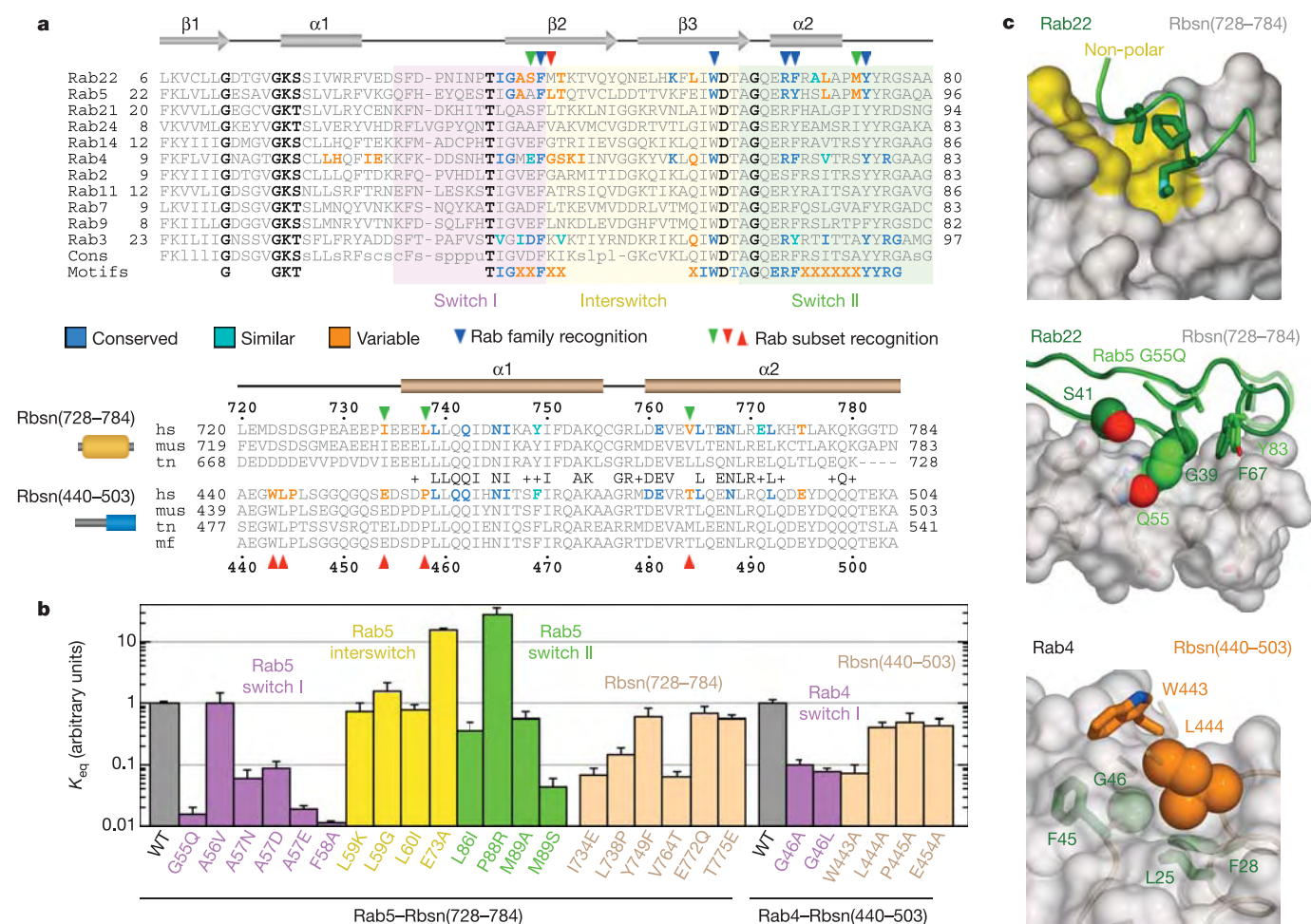


Figure 4 | Structure-based mutational analysis of Rab and rabenosyn-5 interaction specificity. **a**, Local alignment of representative Rab GTPases and rabenosyn-5 homologues. Residues within the binding interfaces for Rab22-Rbsn(728-784), Rab4-Rbsn(440-503), Rab5-rabaptin-5 (ref. 22) and Rab3-rabphilin²¹ are highlighted according to conservation. hs, *Homo sapiens*; mus, *Mus musculus*; tn, *Tetraodon nigroviridis* (puffer fish); mf, *Macaca fascicularis* (macaque). **b**, Effect of site-specific substitutions on

the interaction between Rab5 and Rbsn(728-784) or Rab4 and Rbsn(440-503). Mean values and standard deviations were calculated from two experiments. **c**, Views of the Rab22-Rbsn(728-784) and Rab4-Rbsn(440-503) interfaces relevant to the mutations discussed in the text. Also shown (middle panel) is the structure of the Rab5(G55Q) mutant after superposition with Rab22.

minor effects, with the exception of W443A, which results in a tenfold decrease in affinity. Trp 443 packs against non-polar residues in Rab4, including the invariant Phe 45 in switch I. To facilitate this interaction and the main-chain hydrogen bonds between the β1 strand of Rbsn(440-503) and the β2 strand of Rab4, the side chain of Leu 444 must pack against the Cα of Gly 46. Order of magnitude defects for the G46A and G46L mutations, combined with the lack of a significant defect for the L444A substitution, implicate Leu 444 as a negative determinant that selects against the majority of Rab GTPases.

Consistent with the preceding observations, a triple mutation (E454I, P458L, T484V) that converts predicted specificity determinants in Rbsn(418-503) to the corresponding residues in Rbsn(728-784) exhibits reversed specificity with a preference for Rab5, Rab22 and Rab24 (Supplementary Fig. 7). Eliminating residues 418-447, which are required for binding to Rab4 and Rab14, in the triple mutant has no effect on its affinity for Rab5 (data not shown). A chimera in which the N terminus of Rbsn(418-503) is fused to the helical hairpin of Rbsn(728-784) exhibits a comparable reversal of specificity, resulting in a preference for Rab4 and Rab14, which is further enhanced by an amino acid substitution (V764T) in the helical hairpin.

Comparison with the Rab5-rabaptin²², Rab3-rabphilin²¹ and other GTPase-effector complexes²⁹ reveals similarities in the overall

modes of interaction, which are particularly striking for the effectors containing helical hairpins or coiled coils as the core GTPase binding element³⁰. However, even structurally homologous domains can achieve selective recognition of distinct Rab GTPase subsets, and further discriminate between Rab GTPases with similar active conformations, on the basis of interactions with the switch and interswitch regions. The exquisite specificity of the rabenosyn-5 Rab binding domains is determined by structural as well as compositional diversity. This combination of factors enhances the affinity for Rab-effector subsets with complementary interfaces at the expense of combinations with incompatible structures and/or compositions. The family-wide nature of the recognition process is underscored by the conservation of positive determinants in interacting subsets and negative determinants in non-interacting Rab GTPases. In this respect, the encoding of Rab-effector recognition determinants parallels that observed for the determinants of functional specificity in small GTPases²⁶.

METHODS

Constructs. Constructs were amplified and subcloned into pGEX vectors (Amersham Biosciences) for expression as glutathione S-transferase (GST) fusions or into modified pET vectors (Supplementary Table 1). The modified pET vectors incorporate an N-terminal 6 × His tag with (MGHHHHHHGSLVPRGS) or without (MGHHHHHHGS) a thrombin cleavage site. Mutations were

generated with the Quick Change kit (Stratagene). All constructs were verified by sequencing the entire coding region.

Expression and purification. BL21(DE3) Codon Plus-RIL cells (Stratagene) transformed with expression plasmids were cultured in $2 \times$ YT containing 100 mg l^{-1} ampicillin. For co-expression of modified pET15b-Rbsn(728–784) and modified pET28a-Rab22(Q64L), the cultures were supplemented with 50 mg l^{-1} kanamycin. Cells were grown at 22°C to an optical density at 600 nm of 0.4, induced with 0.05 mM IPTG for 16 h, and disrupted by sonication in 50 mM Tris, pH 7.5 or pH 8.5, 150 mM NaCl, 2 mM MgCl_2 , 0.1% 2-mercaptoethanol, 0.1 mM PMSF and 0.2 mg ml^{-1} lysozyme. After supplementing with 0.5% Triton X-100, the lysates were centrifuged at 35,000g for 40 min. For 6 $\times$ His fusions, supernatants were loaded onto Ni^{2+} -NTA agarose (Qiagen), washed with 50 mM Tris, pH 7.5 or pH 8.5, 500 mM NaCl, 2 mM MgCl_2 , 15 mM imidazole, 0.1% 2-mercaptoethanol and eluted with a gradient of 15–300 mM imidazole. For GST fusions, the supernatants were loaded onto glutathione-Sepharose (Amersham Biosciences), washed with 50 mM Tris, pH 7.5 or pH 8.5, 150 mM NaCl, 2 mM MgCl_2 , 0.1% 2-mercaptoethanol and eluted with 10 mM reduced glutathione. Fusion tags were removed by digestion with thrombin (Haematologic Technologies) or PreScission protease (Amersham Biosciences) and the cleaved proteins isolated over glutathione-Sepharose or Ni^{2+} -NTA agarose. Constructs were further purified by ion exchange over Source Q or Source S followed by gel filtration over Superdex-75 (Amersham Biosciences). Complexes were prepared by co-purification following co-expression (Rab22(Q64L)–Rbsn(728–784)) or by mixing in a 1:1 molar ratio followed by gel filtration over Superdex-75 (Rab4(Q67L)–Rbsn(440–503)).

Nucleotide exchange. Rab GTPases at $2\text{--}10 \text{ mg ml}^{-1}$ were incubated for 6 h with a 25-fold excess of GppNHP in 25 mM Tris, pH 7.5 or pH 8.5, 150 mM NaCl, 5 mM EDTA, 0.1% 2-mercaptoethanol and 100 units per μmol GTPase of calf intestinal alkaline phosphatase (New England Biolabs). After supplementing with 10 mM MgCl_2 , excess nucleotide was removed by gel filtration over Superdex-75 ($> 2 \text{ mg GTPase}$) or a Pierce D-Salt column ($< 2 \text{ mg GTPase}$).

Surface plasmon resonance. Surface plasmon resonance experiments were performed on a BiAcore X instrument. Anti-GST was coupled to CM5 sensor chips using reagents and protocols supplied by the manufacturer. All proteins were dialysed into 10 mM Tris, pH 7.5, 150 mM NaCl, 2 mM MgCl_2 , 0.005% Tween 20 and centrifuged at 1,300g. Sample and reference flow cells were loaded with 800 nM GST fusion or GST, respectively. A flow rate of 0.02 ml min^{-1} was used for all injections. Binding and dissociation were monitored after injection of 6 $\times$ His fusions at varying concentration. After curve alignment and subtraction of the reference sensogram, the equilibrium signal (R_{eq}) at each concentration was extracted by fitting to a Langmuir binding model. Dissociation constants (K_d) were obtained from a fit to $R_{\text{eq}} = R_{\text{max}}/[6 \times \text{His fusion}]/(K_d + [6 \times \text{His fusion}])$.

Sedimentation equilibrium. 6 $\times$ His rabenosyn-5 constructs were dialysed against 50 mM Tris, pH 7.5, 100 mM NaCl and centrifuged to equilibrium in a Beckman Optima XLI. The absorbance at 230 or 280 nm was measured as a function of the radial distance (r) from the axis of rotation. Data were analysed by fitting with the function $A(r) = C_o + C_i \exp(-n\sigma_m(r_o^2 - r^2)/2)$, where C_o and C_i are constants, n represents the oligomeric state, r_o is last data point, and σ_m is calculated with SEDINTERP using the monomer molecular mass.

Crystallization and structure determination. Individual proteins and complexes were crystallized in hanging drops with or without microseeding. Crystals were transferred to a cryostabilizer solution, flash frozen and maintained at 100 K in a nitrogen cryostream (Oxford Cryosystems). Diffraction data were collected at NSLS beamline X25 (complexes and Rab2) and on Rigaku RUH3R generators equipped with Osmic mirrors and Mar30 cm (Mar Research) or R-axis IV (Rigaku) detectors. Data were processed with Denzo and scaled with Scalepack. Structures were solved by molecular replacement using AMoRe, Molrep or Phaser. Crystallographic models were refined by simulated annealing in CNS, automated atom updating with Arp/wArp, minimization with Refmac5, and model building in O. Except where otherwise noted, all programs were used as implemented in CCP4. Additional information related to the structure determination and refinement is compiled in Supplementary Table 2 and Supplementary Fig. 1. Molecular graphics were rendered with PyMol.

Received 16 January; accepted 11 May 2005.

- Pfeffer, S. R. Rab GTPases: specifying and deciphering organelle identity and function. *Trends Cell Biol.* 11, 487–491 (2001).
- Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nature Rev. Mol. Cell Biol.* 2, 107–117 (2001).
- Sivars, U., Aivazian, D. & Pfeffer, S. R. Yip3 catalyses the dissociation of endosomal Rab-GDI complexes. *Nature* 425, 856–859 (2003).
- Rak, A. et al. Structure of Rab GDP-dissociation inhibitor in complex with

prenylated YPT1 GTPase. *Science* 302, 646–650 (2003).

- Calero, M. et al. Dual prenylation is required for Rab protein localization and function. *Mol. Biol. Cell* 14, 1852–1867 (2003).
- Rak, A. et al. Structure of the Rab7:REP-1 complex: insights into the mechanism of Rab prenylation and choroideremia disease. *Cell* 117, 749–760 (2004).
- Seabra, M. C. & Wasmeier, C. Controlling the location and activation of Rab GTPases. *Curr. Opin. Cell Biol.* 16, 451–457 (2004).
- Pfeffer, S. & Aivazian, D. Targeting Rab GTPases to distinct membrane compartments. *Nature Rev. Mol. Cell Biol.* 5, 886–896 (2004).
- Chavrier, P., Parton, R. G., Hauri, H. P., Simons, K. & Zerial, M. Localization of low molecular weight GTP binding proteins to exocytic and endocytic compartments. *Cell* 62, 317–329 (1990).
- Feng, Y., Press, B. & Wandinger-Ness, A. Rab 7: an important regulator of late endocytic membrane traffic. *J. Cell Biol.* 131, 1435–1452 (1995).
- Lombardi, D. et al. Rab9 functions in transport between late endosomes and the trans Golgi network. *EMBO J.* 12, 677–682 (1993).
- Mesa, R., Salomon, C., Roggero, M., Stahl, P. D. & Mayorga, L. S. Rab22a affects the morphology and function of the endocytic pathway. *J. Cell Sci.* 114, 4041–4049 (2001).
- Kauppi, M. et al. The small GTPase Rab22 interacts with EEA1 and controls endosomal membrane trafficking. *J. Cell Sci.* 115, 899–911 (2002).
- Juntula, J. R. et al. Rab14 is involved in membrane trafficking between the Golgi complex and endosomes. *Mol. Biol. Cell* 15, 2218–2229 (2004).
- Simpson, J. C. et al. A role for the small GTPase Rab21 in the early endocytic pathway. *J. Cell Sci.* 117, 6297–6311 (2004).
- Vitale, G. et al. Distinct Rab-binding domains mediate the interaction of Rabaptin-5 with GTP-bound Rab4 and Rab5. *EMBO J.* 17, 1941–1951 (1998).
- de Renzis, S., Sonnichsen, B. & Zerial, M. Divalent Rab effectors regulate the sub-compartmental organization and sorting of early endosomes. *Nature Cell Biol.* 4, 124–133 (2002).
- Fouraux, M. A. et al. Rabip4' is an effector of rab5 and rab4 and regulates transport through early endosomes. *Mol. Biol. Cell* 15, 611–624 (2004).
- Fukuda, M. Distinct Rab binding specificity of Rim1, Rim2, rabphilin, and Noc2. Identification of a critical determinant of Rab3A/Rab27A recognition by Rim2. *J. Biol. Chem.* 278, 15373–15380 (2003).
- Pfeffer, S. R. Structural clues to Rab GTPase functional diversity. *J. Biol. Chem.* 280, 15485–15488 (2005).
- Ostermeier, C. & Brunger, A. T. Structural basis of Rab effector specificity: crystal structure of the small G protein Rab3A complexed with the effector domain of rabphilin-3A. *Cell* 96, 363–374 (1999).
- Zhu, G. et al. Structural basis of Rab5-Rabaptin5 interaction in endocytosis. *Nature Struct. Mol. Biol.* 11, 975–983 (2004).
- Pereira-Leal, J. B. & Seabra, M. C. The mammalian Rab family of small GTPases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the Ras superfamily. *J. Mol. Biol.* 301, 1077–1087 (2000).
- Pereira-Leal, J. B. & Seabra, M. C. Evolution of the Rab family of small GTP-binding proteins. *J. Mol. Biol.* 313, 889–901 (2001).
- Merithew, E. et al. Structural plasticity of an invariant hydrophobic triad in the switch regions of Rab GTPases is a determinant of effector recognition. *J. Biol. Chem.* 276, 13982–13988 (2001).
- Heo, W. D. & Meyer, T. Switch-of-function mutants based on morphology classification of Ras superfamily small GTPases. *Cell* 113, 315–328 (2003).
- Chen, D., Guo, J., Miki, T., Tachibana, M. & Gahl, W. A. Molecular cloning of two novel rab genes from human melanocytes. *Gene* 174, 129–134 (1996).
- Rodriguez-Gabin, A. G., Cammer, M., Almazan, G., Charron, M. & Larocca, J. N. Role of rRAB22b, an oligodendrocyte protein, in regulation of transport of vesicles from trans Golgi to endocytic compartments. *J. Neurosci. Res.* 66, 1149–1160 (2001).
- Vetter, I. R. & Wittinghofer, A. The guanine nucleotide-binding switch in three dimensions. *Science* 294, 1299–1304 (2001).
- Panic, B., Perisic, O., Veprintsev, D. B., Williams, R. L. & Munro, S. Structural basis for Arl1-dependent targeting of homodimeric GRIP domains to the Golgi apparatus. *Mol. Cell* 12, 863–874 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to M. Zerial for a full-length clone of rabenosyn-5; E. Kittler and M. Zapp for assistance with surface plasmon resonance experiments; and A. Delprato for Rab5 mutants. Surface plasmon resonance data were collected in the UMASS Center for AIDS Research Molecular Biology Core. This work was supported by an NIH grant.

Author Information Coordinates and structure factors have been deposited in the Protein Data Bank under the codes 1YZM (Rbsn(458–503)), 1Z0J (Rab22–Rbsn(728–784)), 1Z0K (Rab4–Rbsn(440–503)) and as listed in Supplementary Table 3 (Rab GTPases). Reprints and permissions information is available at ngp.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.G.L. (David.Lambright@umassmed.edu).

LETTERS

Chloride/proton antiporter activity of mammalian CLC proteins CLC-4 and CLC-5

Alessandra Picollo¹ & Michael Pusch¹

CLC-4 and CLC-5 are members of the CLC gene family¹, with CLC-5 mutated in Dent's disease², a nephropathy associated with low-molecular-mass proteinuria and eventual renal failure. CLC-5 has been proposed to be an electrically shunting Cl^- channel in early endosomes, facilitating intraluminal acidification^{3,4}. Motivated by the discovery that certain bacterial CLC proteins are secondary active Cl^-/H^+ antiporters⁵, we hypothesized that mammalian CLC proteins might not be classical Cl^- ion channels but might exhibit Cl^- -coupled proton transport activity. Here we report that CLC-4 and CLC-5 carry a substantial amount of protons across the plasma membrane when activated by positive voltages, as revealed by measurements of pH close to the cell surface. Both proteins are able to extrude protons against their electrochemical gradient, demonstrating secondary active transport. H^+ , but not Cl^- , transport was abolished when a pore glutamate was mutated to alanine (E211A). CLC-0, CLC-2 and CLC-Ka proteins showed no significant proton transport. The muscle channel CLC-1 exhibited a small H^+ transport that might be physiologically relevant. For CLC-5, we estimated that Cl^- and H^+ transport contribute about equally to the total charge movement, raising the possibility that the coupled Cl^-/H^+ transport of CLC-4 and CLC-5 is of significant magnitude *in vivo*.

CLC-5 belongs to a sub-branch of CLC proteins that includes CLC-3 and CLC-4 (ref. 1). These proteins have similar functional properties and are preferentially expressed in intracellular organelles^{1,6–8}. By analogy to other CLC channels^{9,10} they have been assumed to be Cl^- channels as well; however, no clear single-channel activity has yet been reported. CLC-4 (ref. 11) and CLC-5 (A. Accardi and M.P., unpublished results) have a sub-picoSiemens apparent single-channel conductance, an order of magnitude that is compatible with a transporter mechanism. We thus hypothesized that they may have functional similarity to the bacterial CLC-ec1 protein, which is a Cl^-/H^+ antiporter⁵. Strongly outwardly rectifying currents, but no inward currents, can be measured in CLC-5-expressing oocytes⁶. Therefore, a possible contribution of proton movement to the measured current can not be assessed using reversal potential measurements, as has been done for CLC-ec1 (ref. 5). We therefore used pH-sensitive microelectrodes¹² to detect H^+ movement. Applying trains of positive voltage pulses and placing the pH-sensitive microelectrode close to CLC-5-expressing oocytes, robust changes in extracellular pH could be recorded (Fig. 1). The pH change correlated with the voltage activation of CLC-5 (Fig. 1a). Applying the voltage clamp leads immediately to an acidification from pH 7.4 to almost 7.1 within about 90 s for the experiment shown in Fig. 1a, b. Switching off the voltage clamp (arrows in Fig. 1) leads to an almost full recovery within about 60 s (Fig. 1b). The immediate onset of acidification was seen in all CLC-5-expressing oocytes, with time constants ranging from 7 to 30 s. The low concentration of 0.5 mM HEPES is critical, as experiments with 5 mM HEPES yielded only a very small response (data not shown). There was no clear correlation

of pH change with the current expression of different oocytes (data not shown), probably due to the variability of the exact positioning of the pH-sensitive microelectrode. However, within the same oocyte more positive test voltages (V_p) and/or longer pulses yielded larger changes in pH (Fig. 1c).

In the absence of extracellular Cl^- no pH change could be detected (Fig. 1d). This shows that H^+ transport through CLC-5 depends on chloride and is not mediated by a passive, independent pathway. For Cl^-/H^+ antiport, alkalization might be expected under zero Cl^- conditions, as those used in Fig. 1d, but the strong outward rectification of CLC-5 probably renders inwardly directed H^+ transport inefficient. To test whether H^+ transport is directly coupled to Cl^- movement, we imposed an inwardly directed proton gradient, bathing the oocyte in a solution with pH 5.8. Under these conditions, intracellular pH remains stable, above 7.2, as measured with intracellular pH-sensitive microelectrodes in separate oocytes (data not shown). Thus, the H^+ equilibrium potential is $E_H > 80$ mV. For voltages $< E_H$, net proton movement through any passive, diffusive H^+ transport mechanism can only be inwardly directed. Yet, activation of CLC-5 at 60 mV led to a robust acidification (Fig. 1e). This demonstrates that the energy of the downhill Cl^- movement is used to actively extrude protons.

We performed several controls to assure that the pH response is specific for CLC-5. No pH response was observed in non-injected oocytes (Fig. 2a). We next investigated the mutation E211A. The conserved glutamate is important for the gating of CLC-0 and other CLC channels^{13–16}, and in CLC-ec1 the corresponding E148A mutation abolishes H^+ transport⁵. In CLC-5 the E211A mutation leads to a loss of the extreme outward rectification¹³ (Fig. 2b, e). Notably, H^+ transport was completely abolished by the mutation (Fig. 2b). In addition, the mutation led to a loss of inhibition of currents at low extracellular pH (Fig. 2c–f). Reduced currents of CLC-5 at low pH are consistent with a Cl^-/H^+ antiporter because the net driving force is reduced, and the lack of inhibition of E211A is in accordance with its inability to transport protons.

Although the overall phenotype of CLC-3 is very similar to that of CLC-5 (ref. 8; see also Fig. 3a) it does not reproducibly express currents in oocytes, and we were thus unable to investigate a possible proton transport activity. CLC-4 expresses currents well, is similar to CLC-5 (ref. 13; see also Fig. 3b, inset), and its activation was associated with robust H^+ transport (Fig. 3b). Similar to CLC-5, CLC-4 was also able to transport protons against their electrochemical gradient (Fig. 3c).

The human proteins CLC-1, CLC-2, CLC-Ka and CLC-Kb, and the *Torpedo* channel CLC-0, are clearly associated with ion channel activity^{9,10,17–19} (A.P. and M.P., unpublished results). In fact, no H^+ transport could be detected for CLC-0, CLC-2($\Delta 16–61$) and CLC-Ka (Fig. 3d–f). Surprisingly, a small, but reproducible change in pH was detected for CLC-1 (Fig. 3g, h).

We determined the relative contribution of H^+ transport to CLC-5

¹Istituto di Biofisica, Consiglio Nazionale delle Ricerche, Via de Marini 6, I-16149 Genova, Italy.

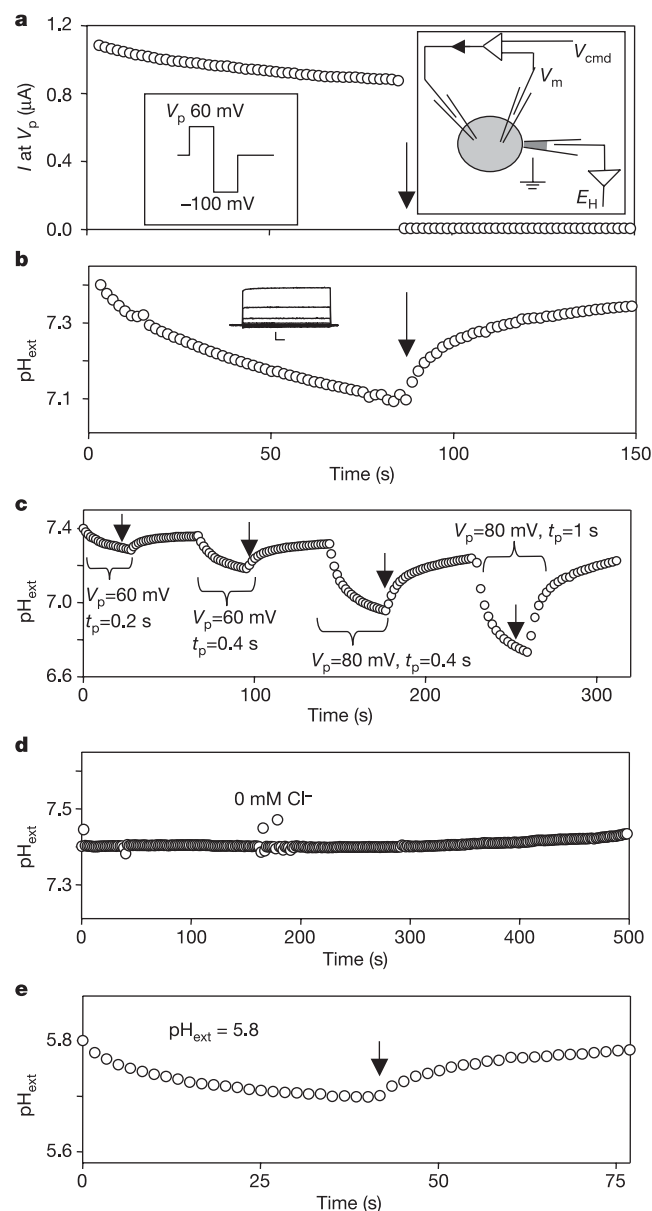


Figure 1 | Acidification mediated by CIC-5. **a**, **b**, A representative measurement of extracellular pH (pH_{ext}) close to the oocyte is shown. The right inset illustrates the measurement configuration (V_m , membrane potential; V_{cmd} , command potential; E_H , potential recorded by the pH-sensitive microelectrode). Pulses to voltage V_p (left inset) elicit a large current (shown in **a** as a function of time), whereas the response to pulses to -100 mV is mostly endogenous. These 'leak' pulses were included to compensate for unspecific proton conductance. The arrow in **b** (and in all other figures) indicates switch off of the voltage clamp. The inset in **b** shows a family of voltage-clamp traces evoked by pulses from -140 to 80 mV (scale bars: $0.5 \mu A$, 5 ms). **c**, The parameters of the pulse protocol (t_p , duration of the voltage pulse to V_p) were varied as indicated. Time periods of voltage clamp at the indicated voltages are depicted by brackets. **d**, Results in a Cl^- free solution. The same oocyte produced a robust change in extracellular pH in the standard Cl^- solution ($n = 4$ oocytes tested). **e**, Experiment in a pH 5.8 solution ($n = 5$ oocytes tested).

currents by measuring acidification of a small volume achieved by single oocytes. The charge of the transported protons, Q_H , was determined from the change in pH (see Methods for description). The total transported charge, Q_T , was calculated from the integral of the current. Values for Q_T and Q_H for individual experiments are shown in Table 1. The values for Q_T and Q_H are of the same order of

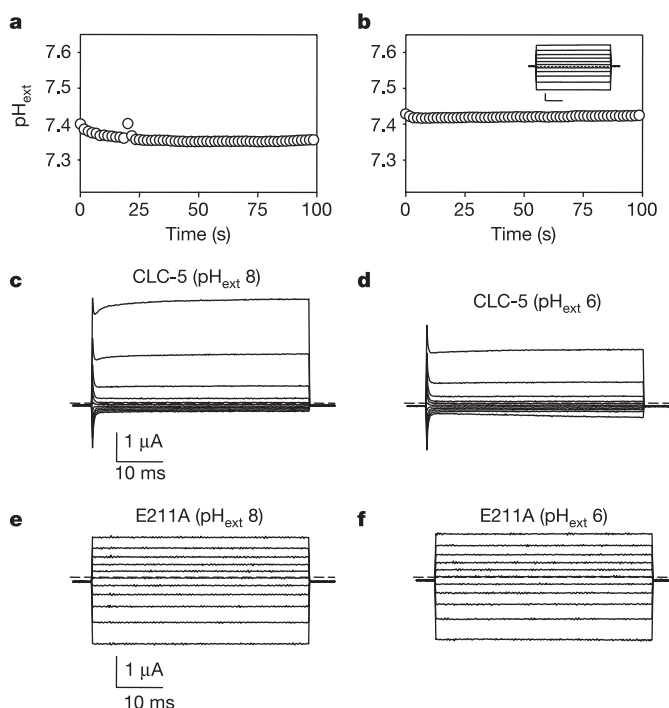


Figure 2 | Proton transport is specific for CIC-5 and is abolished by a pore mutation. **a**, **b**, Control recordings with a non-injected oocyte (**a**) and an oocyte injected with CIC-5 E211A (**b**) (no leak pulses were applied in these measurements) ($n \geq 3$ each). The inset in **b** shows voltage-clamp traces from the same oocyte evoked by pulses from -140 to 80 mV (scale bars: $5 \mu A$, 10 ms). **c**–**f**, Voltage-clamp traces evoked by steps from 80 mV to -140 mV from an oocyte expressing wild-type CIC-5 (**c**, **d**) or its mutant E211A (**e**, **f**) at the indicated extracellular pH values ($n = 4$ each).

magnitude for each experiment, demonstrating that proton movement is responsible for a major component of the measured currents. The ratio $r = Q_H/Q_T$ is quite variable, but excluding values $r > 1$, which are not compatible with a Cl^-/H^+ antiport mechanism (indicated by an asterisk in Table 1), we obtain an estimate of transport stoichiometry of 1.1 ± 0.5 for H^+ versus Cl^- . Considering the relatively large error associated with the pH measurement, this value represents only a rough estimate compared to the more precise equilibrium measurement of ref. 5 for CIC-ec1: a value of 0.5 ($H^+:Cl^-$). The two values are, nevertheless, surprisingly similar. The roughly equal contribution of Cl^- and H^+ movement implies that H^+ transport mediated by CIC-5 is not just a relict of irrelevant magnitude but is significant.

On the basis of structural and functional similarity with CIC-4 and CIC-5, CIC-3 is also a likely Cl^-/H^+ antiporter, but, as expected, we found no significant H^+ transport for the plasma membrane channels CIC-0, CIC-2($\Delta 16$ – 61) and CIC-Ka at expression levels comparable to those achieved with CIC-5. Any residual H^+ transport is probably negligible compared with the channel-mediated Cl^- transport. It remains to be determined whether the small H^+ transport signal seen for CIC-1 is of physiological relevance for its presumed role in stabilizing the muscle membrane potential.

The proteins CIC-3, CIC-4 and CIC-5 have been assumed to provide a shunt conductance in the membranes of intracellular organelles, permitting an efficient acidification by a V-type H^+ -ATPase¹. However, their strong outward rectification together with a presumed slightly positive inside luminal potential²⁰ makes this hypothesis unlikely. Furthermore, the Cl^-/H^+ antiport function found here renders the interpretation of the physiological role even more complicated. Nevertheless, the comparison of acidification rates from wild-type and knockout mice supports a role in vesicle acidification^{7,21,22}. It may thus be that the low activity of CIC-3, CIC-4

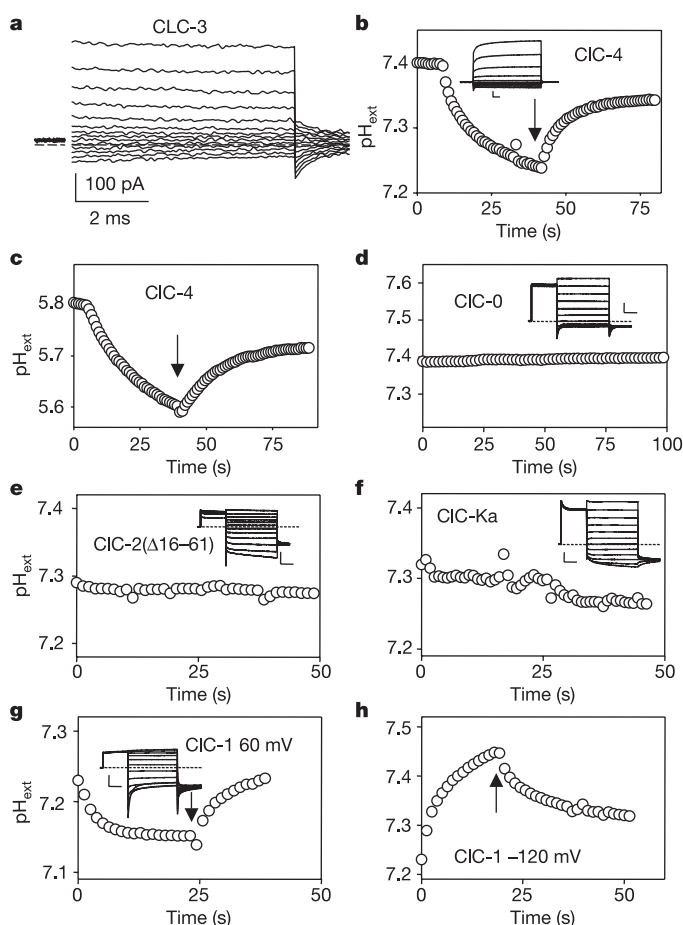


Figure 3 | H^+ transport of other CLC proteins. **a**, Patch-clamp traces of CLC-3 produced by steps from 160 mV to -140 mV. Initial capacity transient is blanked. **b–h**, Extracellular pH close to oocytes expressing the indicated channel types bathed in standard solution with 0.5 mM HEPES (except panel **c**, the solution for which contained 0.5 mM MES, pH 5.8) after activation by 0.4-s pulses to 60 mV (except for **h**, for which $V_p = -120$ mV) ($n \geq 5$ for each channel type). No leak pulses were applied for CLC-0, CLC-2($\Delta 16-61$), CLC-Ka and CLC-1. Insets show families of voltage-clamp traces evoked in the respective oocytes by pulses from -140 to 80 mV (scale bars: 0.5 μ A, 5 ms (**b**); 2 μ A, 50 ms (**d**); 1 μ A, 50 ms (**e**); 2 μ A, 50 ms (**f**); 2 μ A, 50 ms (**g**)).

and CLC-5 at physiological membrane voltages is sufficient to provide enough shunt for some acidification to occur. The Cl^-/H^+ antiport would actually lead to a 'waste' of accumulated protons and could constitute part of a postulated significant H^+ leak²³. It might even be speculated that the unfavourable conditions for CLC-5 activation constitute a brake to limit acidification to a certain degree, the one optimal for endosome function. An alternative, more speculative role of CLC-3, CLC-4 and CLC-5 might be in the fusion of intracellular organelles. It becomes increasingly evident that the fusion of various vesicles and organelles requires an elevated local calcium concentration produced by release from the endosomes themselves^{24,25}. The Ca^{2+} -release pathway is unknown, but if release is electrogenic it would lead to a negative inside luminal potential that could activate CLC-5. The Ca^{2+} release could thereby lead to CLC-5-mediated acidification and, furthermore, CLC-5-mediated charge compensation could be essential for an efficient fusion process. This could indirectly explain the above-mentioned results from knockout studies^{7,21,22}, because endosomes from CLC-5 knockout mice might not mature properly into an acidification-competent state. Finally, there remains the possibility that accessory proteins are necessary for

Table 1 | Contribution of protons to total charge transport

Total charge (Q_T ; μ C)	Proton charge (Q_H ; μ C)	Ratio (Q_H/Q_T)
30.7	23.8	0.78
33.7	3.7	0.11
18.5	3.9	0.21
11.8	24.4	2.07*
14.6	13.6	0.93
9.6	23.8	2.47*
11.0	23.7	2.17*
16.0	10.4	0.65
13.7	6.5	0.47

Each row shows the result from an individual oocyte. The first column shows the total charge transfer obtained by integrating the current signal. The second column shows the charge of the transferred protons obtained from the measured bulk pH change as described in Methods. The third column shows the ratio of the proton charge and the total charge.

* Values >1 , which are not compatible with a Cl^-/H^+ antiport mechanism.

the proper Cl^-/H^+ antiporter function of CLC-3, CLC-4 and CLC-5. Whatever the precise role of CLC-3, CLC-4 and CLC-5 turns out to be, it will be that of a secondary active ion transporter and not an ion channel.

METHODS

Complementary DNA constructs and oocyte injection. Human CLC-4 and CLC-5 were in the vector pFrog3. Human CLC-3 was in the pCI vector. CLC-0, human CLC-1, rat CLC-2($\Delta 16-61$) and human CLC-Ka were in the pTLN vector²⁶. CLC-0 contained the C212S mutation known to remove the slow inactivation²⁷ and CLC-2 contained an amino-terminal deletion ($\Delta 16-61$) resulting in much larger currents in oocytes^{15,28}. CLC-Ka was co-expressed with barttin²⁹. The CLC-5 E211A mutation was introduced using recombinant polymerase chain reaction. cRNA was generated and *Xenopus* oocytes were obtained, injected and incubated as described¹⁵. Human CLC-3 was transiently transfected in tsA201 cells and analysed using the patch clamp technique. The bath solution contained (in mM): 150 NMDG-Cl, 10 HEPES, 1.8 $CaCl_2$, 1 $MgCl_2$, pH 7.3, whereas the pipette solution contained 130 NMDG-Cl, 2 EGTA, 10 HEPES, 1 $MgCl_2$, pH 7.3. Positively transfected cells were identified by co-transfection of CD8 (ref. 30).

Voltage clamp and pH measurements. A Tec-03x amplifier (npi electronic) was used for voltage clamp, and data were acquired with a custom program (GePulse). Pulse protocols are described in the figure legends. pH-sensitive microelectrodes were pulled identical to the voltage-clamp pipettes, silanized with dichlorodimethylsilane (Sigma), backfilled with the proton ionophore B (Fluka) and then filled with a solution containing (in mM): 150 NaCl, 23 NaOH, 40 KH_2PO_4 , pH 6.8. Pipettes responded with a slope of 57–63 mV per pH unit. They were connected to a custom-built electrometer with an input impedance of $>10^{15}$ Ohm. The bath solution for pH measurements contained (in mM): 100 NaCl, 5 $MgCl_2$, 0.5 HEPES, pH 7.4. Cl^- free solution (Fig. 1d) contained (in mM): 100 Na-glutamate, 5 $MgSO_4$, 0.5 HEPES, pH 7.4. The pH 5.8 solution (Fig. 1e) contained (in mM): 100 NaCl, 5 $MgCl_2$, 0.5 MES, pH 5.8.

For the determination of the amount of transported protons, oocytes were placed in 20 or 30 μ l solution containing (in mM): 100 NaCl, 5 $MgCl_2$, 0.1 HEPES, 0.1 amiloride. pH was measured before and after voltage-clamp stimulation by 20–60 0.4-s pulses of the type shown in Fig. 1a (inset). Solution was mixed before pH measurements. From the pH change, the amount of transferred protons was determined from a separately determined titration curve of the solution. A typical pH change of 0.2 corresponded to an added amount of 8 μ M (in 20 μ l) and thus to a transferred charge of 16 μ C. Values of pH change measured from non-injected oocytes (mean value of 0.05) were much smaller than those from CLC-5-expressing oocytes (mean value of 0.18), and were subtracted.

Received 10 February; accepted 4 May 2005.

- Jentsch, T. J., Poet, M., Fuhrmann, J. C. & Zdebek, A. A. Physiological functions of CLC Cl channels gleaned from human genetic disease and mouse models. *Annu. Rev. Physiol.* **67**, 779–807 (2005).
- Lloyd, S. E. *et al.* A common molecular basis for three inherited kidney stone diseases. *Nature* **379**, 445–449 (1996).
- Piwon, N., Günther, W., Schwake, M., Bösl, M. R. & Jentsch, T. J. CLC-5 Cl^- -channel disruption impairs endocytosis in a mouse model for Dent's disease. *Nature* **408**, 369–373 (2000).
- Wang, S. S. *et al.* Mice lacking renal chloride channel, CLC-5, are a model for Dent's disease, a nephrolithiasis disorder associated with defective receptor-mediated endocytosis. *Hum. Mol. Genet.* **9**, 2937–2945 (2000).

5. Accardi, A. & Miller, C. Secondary active transport mediated by a prokaryotic homologue of CIC-Cl⁻ channels. *Nature* **427**, 803–807 (2004).
6. Steinmeyer, K., Schwappach, B., Bens, M., Vandewalle, A. & Jentsch, T. J. Cloning and functional expression of rat CLC-5, a chloride channel related to kidney disease. *J. Biol. Chem.* **270**, 31172–31177 (1995).
7. Stobrawa, S. M. *et al.* Disruption of CIC-3, a chloride channel expressed on synaptic vesicles, leads to a loss of the hippocampus. *Neuron* **29**, 185–196 (2001).
8. Li, X., Shimada, K., Showalter, L. A. & Weinman, S. A. Biophysical properties of CIC-3 differentiate it from swelling-activated chloride channels in Chinese hamster ovary-K1 cells. *J. Biol. Chem.* **275**, 35994–35998 (2000).
9. Miller, C. Open-state substructure of single chloride channels from Torpedo electroplax. *Phil. Trans. R. Soc. Lond. B* **299**, 401–411 (1982).
10. Saviane, C., Conti, F. & Pusch, M. The muscle chloride channel CIC-1 has a double-barreled appearance that is differentially affected in dominant and recessive myotonia. *J. Gen. Physiol.* **113**, 457–468 (1999).
11. Hebeisen, S. *et al.* Anion permeation in human CIC-4 channels. *Biophys. J.* **84**, 2306–2318 (2003).
12. Tsai, T. D., Shuck, M. E., Thompson, D. P., Bienkowski, M. J. & Lee, K. S. Intracellular H⁺ inhibits a cloned rat kidney outer medulla K⁺ channel expressed in *Xenopus* oocytes. *Am. J. Physiol.* **268**, C1173–C1178 (1995).
13. Friedrich, T., Breiderhoff, T. & Jentsch, T. J. Mutational analysis demonstrates that CIC-4 and CIC-5 directly mediate plasma membrane currents. *J. Biol. Chem.* **274**, 896–902 (1999).
14. Dutzler, R., Campbell, E. B. & MacKinnon, R. Gating the selectivity filter in CIC chloride channels. *Science* **300**, 108–112 (2003).
15. Estévez, R., Schroeder, B. C., Accardi, A., Jentsch, T. J. & Pusch, M. Conservation of chloride channel structure revealed by an inhibitor binding site in CIC-1. *Neuron* **38**, 47–59 (2003).
16. Traverso, S., Elia, L. & Pusch, M. Gating competence of constitutively open CLC-0 mutants revealed by the interaction with a small organic inhibitor. *J. Gen. Physiol.* **122**, 295–306 (2003).
17. Bauer, C. K., Steinmeyer, K., Schwarz, J. R. & Jentsch, T. J. Completely functional double-barreled chloride channel expressed from a single Torpedo cDNA. *Proc. Natl Acad. Sci. USA* **88**, 11052–11056 (1991).
18. Weinreich, F. & Jentsch, T. J. Pores formed by single subunits in mixed dimers of different CLC chloride channels. *J. Biol. Chem.* **276**, 2347–2353 (2001).
19. Chen, T.-Y. Structure and function of CLC channels. *Annu. Rev. Physiol.* **67**, 809–839 (2005).
20. Van Dyke, R. W. & Belcher, J. D. Acidification of three types of liver endocytic vesicles: similarities and differences. *Am. J. Physiol. Cell Physiol.* **266**, C81–C94 (1994).
21. Hara-Chikuma, M., Wang, Y., Guggino, S. E., Guggino, W. B. & Verkman, A. S. Impaired acidification in early endosomes of CIC-5 deficient proximal tubule. *Biochem. Biophys. Res. Commun.* **329**, 941–946 (2005).
22. Hara-Chikuma, M. *et al.* CIC-3 chloride channels facilitate endosomal acidification and chloride accumulation. *J. Biol. Chem.* **280**, 1241–1247 (2005).
23. Chandy, G., Grabe, M., Moore, H.-P. H. & Machen, T. E. Proton leak and CFTR in regulation of Golgi pH in respiratory epithelial cells. *Am. J. Physiol. Cell Physiol.* **281**, C908–C921 (2001).
24. Holroyd, C., Kistner, U., Annaert, W. & Jahn, R. Fusion of endosomes involved in synaptic vesicle recycling. *Mol. Biol. Cell* **10**, 3035–3044 (1999).
25. Bayer, M. J., Reese, C., Buhler, S., Peters, C. & Mayer, A. Vacuole membrane fusion: V0 functions after trans-SNARE pairing and is coupled to the Ca²⁺-releasing channel. *J. Cell Biol.* **162**, 211–222 (2003).
26. Lorenz, C., Pusch, M. & Jentsch, T. J. Heteromultimeric CLC chloride channels with novel properties. *Proc. Natl Acad. Sci. USA* **93**, 13362–13366 (1996).
27. Lin, Y. W., Lin, C. W. & Chen, T. Y. Elimination of the slow gating of CIC-0 chloride channel by a point mutation. *J. Gen. Physiol.* **114**, 1–12 (1999).
28. Gründer, S., Thiemann, A., Pusch, M. & Jentsch, T. J. Regions involved in the opening of CIC-2 chloride channel by voltage and cell volume. *Nature* **360**, 759–762 (1992).
29. Estévez, R. *et al.* Barttin is a Cl⁻ channel β -subunit crucial for renal Cl⁻ reabsorption and inner ear K⁺ secretion. *Nature* **414**, 558–561 (2001).
30. Jurman, M. E., Boland, L. M., Liu, Y. & Yellen, G. Visual identification of individual transfected cells for electrophysiology using antibody-coated beads. *Biotechniques* **17**, 876–881 (1994).

Acknowledgements We thank L. Elia for technical assistance, T. Jentsch and A. Zdebik for providing the CIC-0, CIC-1, CIC-2(Δ 16–61), CIC-3, CIC-4, CIC-Ka, barttin and CIC-5 cDNAs, E. Gaggero and G. Gaggero for constructing the high-impedance amplifier, and G. Gaggero for help in constructing the measuring chamber. Financial support by Telethon Italy and the Italian Research Ministry is gratefully acknowledged.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.P. (pusch@ge.ibf.cnr.it).

LETTERS

Voltage-dependent electrogenic chloride/proton exchange by endosomal CLC proteins

Olaf Scheel^{1*}, Anselm A. Zdebik^{1*}, Stéphane Lourdé¹ & Thomas J. Jentsch¹

Eukaryotic members of the CLC gene family function as plasma membrane chloride channels, or may provide neutralizing anion currents for V-type H⁺-ATPases that acidify compartments of the endosomal/lysosomal pathway¹. Loss-of-function mutations in the endosomal protein CLC-5 impair renal endocytosis² and lead to kidney stones³, whereas loss of function of the endosomal/lysosomal protein CLC-7 entails osteopetrosis⁴ and lysosomal storage disease⁵. Vesicular CLCs have been thought to be Cl⁻ channels, in particular because CLC-4 and CLC-5 mediate plasma membrane Cl⁻ currents upon heterologous expression^{6,7}. Here we show that these two mainly endosomal CLC proteins instead function as electrogenic Cl⁻/H⁺ exchangers (also called antiporters), resembling the transport activity of the bacterial protein CLC-e1 (ref. 8), the crystal structure of which has already been determined⁹. Neutralization of a critical glutamate residue not only abolished the steep voltage-dependence of transport⁷, but also eliminated the coupling of anion flux to proton counter-transport. CLC-4 and CLC-5 may still compensate the charge accumulation by endosomal proton pumps, but are expected to couple directly vesicular pH gradients to Cl⁻ gradients.

To investigate whether CLC-4 or CLC-5 display Cl⁻/H⁺ exchange activity, as observed for the *Escherichia coli* protein CLC-e1 (ref. 8), we measured the intracellular pH of transfected tsA201 cells using the pH-dependent fluorescence of 2',7'-bis-(2-carboxyethyl)-5-(and-6)-carboxyfluorescein (BCECF). In apparent contrast to such an exchange activity, intracellular pH did not change when lowering the extracellular Cl⁻ concentration ([Cl⁻]_o). However, currents of CLC-4 and CLC-5 were only detected at positive intracellular voltages^{6,7}. If these currents reflect electrogenic Cl⁻/H⁺ exchange, H⁺ transport may only be observed upon depolarization. Intracellular pH was therefore recorded in response to changes of the plasma membrane voltage. The plasma membrane was patch-clamped and exposed to different voltages using the gramicidin-perforated patch technique that prevented the loss of BCECF and limited the exchange of protons with the patch pipette. If there was an electrogenic exchange of Cl⁻ for H⁺, then depolarizing the membrane should lead to an efflux of H⁺; that is, to a cytosolic alkalinization. Indeed, exposure of patch-clamped cells transfected with CLC-4 or CLC-5 to positive voltages resulted in an increase of intracellular pH (Fig. 1a–c). No significant change in intracellular pH was detected when imposing negative voltages (Fig. 1a). The voltage dependence of the rate of intracellular pH change ($\Delta\text{pH}_i/\Delta t$; Fig. 1e, f) correlated well with the steep voltage dependence of CLC-4 or CLC-5 currents^{6,7} (Fig. 2a). Depolarization of cells changes the driving force for H⁺ and may cause alkalinization in the presence of an H⁺ conductance; however, non-transfected cells lacked significant changes to intracellular pH in response to depolarization. Furthermore, depolarizing cells expressing a Cl⁻ conductance will increase intracellular Cl⁻ concentration ([Cl⁻]_i), which might, in turn, cause an alkalinization of the cytoplasm

through endogenous Cl⁻/HCO₃⁻ (or Cl⁻/OH⁻) exchangers. However, control cells expressing the *Torpedo* Cl⁻ channel CLC-0 did not alkalinize upon depolarization, although their Cl⁻ currents were of similar magnitude (Fig. 1d).

Hence, the depolarization-induced alkalinization of cells expressing CLC-4 or CLC-5 was neither due to a changed driving force for H⁺, nor to an intracellular accumulation of chloride. Strong evidence for Cl⁻/H⁺ exchange came from experiments in which H⁺ was driven against its electrochemical gradient (Fig. 1c). Cells expressing CLC-4 or CLC-5 were exposed to an extracellular pH of 5.0, creating a chemical H⁺ gradient of 120 mV (assuming an intracellular pH of 7.0), from which the depolarization to +60 mV must be subtracted to obtain the electrochemical gradient for H⁺. Hence, the resulting cytoplasmic alkalinization (Fig. 1c) demonstrated H⁺ transport against an electrochemical gradient of more than 60 mV.

In the bacterial CLC-e1 protein, the coupling of chloride flux to protons was abolished when a critical glutamate was mutated to alanine (E148A)⁸. In the CLC-e1 crystal, the negative side chain of this glutamate blocks the access of extracellular anions to the narrowest part of the pore⁹. The equivalent mutations E224A and E211A in CLC-4 and CLC-5, respectively, converted their currents from being strongly outwardly rectifying to having a nearly ohmic behaviour⁷, as shown for CLC-5 in Fig. 2a, b. Whereas extracellular acidification reduced the currents of wild-type CLC-4 and CLC-5 proteins⁷, the anion conductance of the respective glutamate mutants was independent of extracellular pH (Fig. 2a–c). This suggests that these mutations, similar to CLC-e1 (ref. 8), eliminate the H⁺ coupling of anion currents. Indeed, depolarizing cells expressing these mutants failed to induce alkalinization (Fig. 2d).

If CLC-4 and CLC-5 are Cl⁻/H⁺ exchangers, changes in extracellular anion concentrations should affect the counter-transport of H⁺. To depolarize cells to voltages compatible with the transport activity of CLC-4 or CLC-5, they were co-transfected with the peptide-gated snail Na⁺ channel FaNaC^{10,11} (Fig. 3). Patch-clamp experiments revealed that its ligand H-Phe-Met-Arg-Phe-NH₂ (FMRFamide) depolarized the cells to roughly +30 mV. Exposure of cells co-expressing CLC-5 and FaNaC to FMRFamide induced alkalinization with 150 mM, but not with 4 mM, extracellular Cl⁻ (Fig. 3a). Such effects were not seen when FaNaC was expressed alone or together with the CLC-5 E211A mutant. The alkalinization observed upon exposure to FMRFamide and 150 mM Cl⁻ could be reversed by lowering [Cl⁻]_o to 4 mM (Fig. 3b), just opposite to pH changes expected with anion exchangers. This intracellular pH recovery was absent when cells were hyperpolarized by Na⁺ removal to inhibit Cl⁻/H⁺ exchange (Fig. 3c), suggesting that Cl⁻/H⁺ exchange mediates the pH recovery.

The observed decrease of currents upon extracellular acidification⁷ (Fig. 2a, c) is compatible with the notion that currents directly reflect an electrogenic Cl⁻/H⁺ exchange, because increasing [H⁺]_o would

¹Zentrum für Molekulare Neurobiologie, ZMNH, Universität Hamburg, Falkenried 94, D-20246 Hamburg, Germany.

*These authors contributed equally to this work.

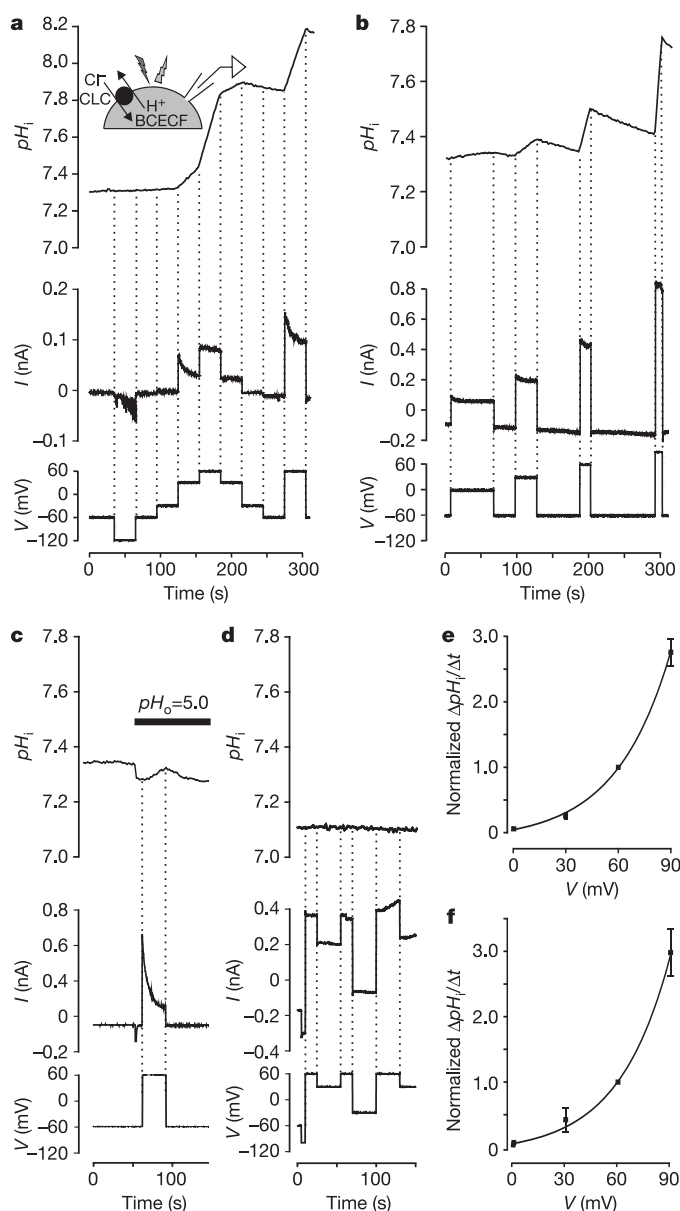


Figure 1 | Depolarization alkalinizes cells expressing CIC-4 or CIC-5, but not those expressing CIC-0. **a–d**, Measurements of intracellular pH (pH_i ; top panel), clamp current (middle panel) and clamp voltage (bottom panel) of tsA201 cells transfected with CIC-4 (**a**), CIC-5 (**b**), or CIC-0 (**c**). Cells were voltage clamped using gramicidin-perforated patches, and intracellular pH was determined using BCECF fluorescence (inset in **a**). Current relaxations with depolarizing pulses may reflect a rise in $[Cl^-]_i$. Similar results were obtained with 6, 13 and 7 cells for CIC-4, CIC-5 and CIC-0, respectively. **c**, Depolarization alkalinizes a CIC-4-transfected cell also with an extracellular pH (pH_o) of 5.0, demonstrating H^+ transport against its electrochemical gradient. Similar results were obtained with 5 and 9 cells for CIC-4 and CIC-5, respectively. Intracellular pH also dropped upon extracellular acidification in untransfected controls. The acidification upon returning to -60 mV (**a–c**) probably represents pH equilibration over the patch. **e, f**, Rates of intracellular pH change as a function of clamp voltage for CIC-4 (**e**) and CIC-5 (**f**). Results were obtained using protocols as in **b**, and represent means from 5 and 8 cells for CIC-4 and CIC-5, respectively. Data were normalized to $\Delta pH_i/\Delta t$ at 60 mV. Error bars indicate s.e.m.

lower the gradient driving the exchanger. Unfortunately, the extreme outward rectification of currents mediated by CIC-3, CIC-4 and CIC-5 precludes a reliable determination of reversal potentials that could be used to calculate the coupling ratio, as done for the bacterial protein CIC-e1 (ref. 8). We therefore estimated the stoichiometry of

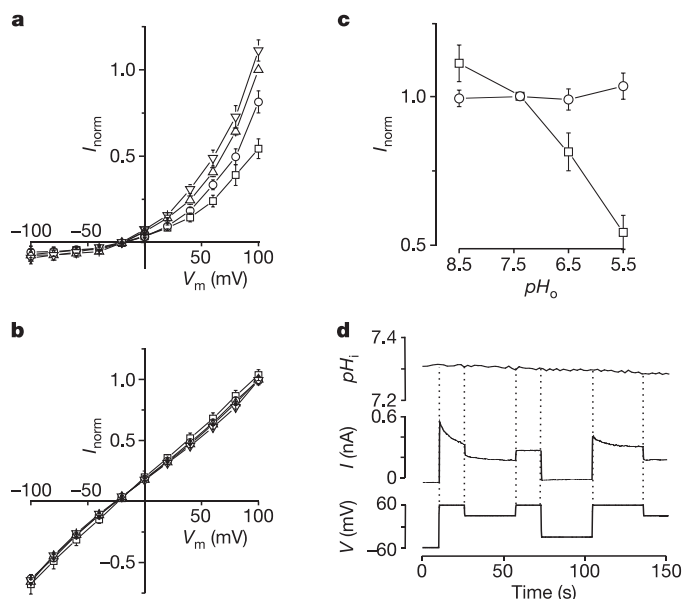


Figure 2 | The E211A mutation abolishes flux coupling to H^+ . **a, b**, Steady-state I - V for CIC-5 (**a**) and CIC-5(E211A) (**b**) expressed in *Xenopus* oocytes. Extracellular pH was varied between pH 5.5 (squares), 6.5 (circles), 7.4 (triangles) and 8.5 (inverted triangles). Currents were normalized for individual oocytes to currents at $+100$ mV in ND96 (pH 7.4). Mean currents at 60 mV were $1.18 \pm 0.17 \mu A$ ($\pm$ s.e.m.) for CIC-5 and $1.07 \pm 0.36 \mu A$ ($\pm$ s.e.m.) for CIC-5(E211A). Averages are from six oocytes (**a**) and ten oocytes (**b**) (two batches each). **c**, Extracellular pH dependence of currents at $+100$ mV for CIC-5 (squares) and CIC-5(E211A) (circles). Currents were normalized to values at pH 7.4. Similar effects were seen for CIC-4(E224A) (data not shown). **d**, Intracellular pH of a tsA201 cell expressing CIC-5(E211A) is unaffected by depolarization (done as in Fig. 1a–c). Similar results were obtained in 11 and 4 experiments for CIC-5(E211A) and CIC-4(E224A), respectively.

Cl^-/H^+ coupling by comparing the rate of alkalinization to currents measured in experiments such as the one shown in Fig. 1b. For individual cells expressing either CIC-4 or CIC-5, their intracellular pH change ($\Delta pH_i/\Delta t$) was used to estimate H^+ fluxes. The calculation was based on the published buffer capacity of HEK cells¹² and on estimates of cell volume obtained from the area occupied by the individual cell (see Methods). These H^+ fluxes were related to clamp currents to yield estimates for the coupling stoichiometry for an exchange of nCl^- for $1H^+$. Values thus obtained for individual cells were averaged. Under these assumptions, n was calculated as 1.6 ± 0.7 (s.d., 6 cells) for CIC-4 and $n = 1.5 \pm 0.7$ (11 cells) for CIC-5. However, uncertainties in buffer capacity^{12,13}, cell volume and possible contributions of leak currents and H^+ transport over the perforated patch¹⁴ only allow us to give a rough estimate of $1 \leq n \leq 5$. This estimate agrees well with the approximate stoichiometry of 2 for CIC-e1, as determined with the easier and more exact determination of reversal potentials⁸. H^+ and Cl^- transport mediated by CIC-4 and CIC-5 have the same order of magnitude, suggesting that H^+ transport by endosomal CLCs might have a functional impact.

In the bacterial protein, flux coupling required the presence of a glutamate in the extracellular access pathway to the pore⁸. Similarly, fluxes of CIC-4 and CIC-5 were not just uncoupled in the absence of this glutamate, but also, similar to CIC-e1, both mammalian proteins displayed a low or negligible permeability to H^+ . Experiments such as the one shown in Fig. 2d impose an upper limit on the H^+ permeability of the glutamate mutant ($<1/20$ of the Cl^- permeability). It appears to behave like a pure Cl^- conductance. Because current magnitudes were not significantly increased by this mutation (Fig. 2) (with comparable protein levels; data not shown), the

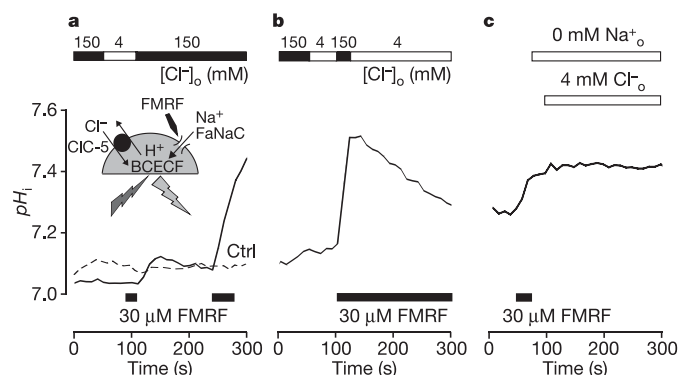


Figure 3 | H^+ transport depends on chloride and voltage. **a**, tsA201 cells transfected with CIC-5 and the ligand-gated Na^+ channel FaNaC¹⁰ were loaded with the pH indicator BCECF (inset). FMRFamide depolarized such cells to 28 ± 4 mV ($n = 8$) in Ringer's solution (data not shown) and changed intracellular pH with 150 mM, but not with 4 mM, Cl^- . FaNaC remained partially activated when removing FMRFamide, resulting in a moderate alkalinization with 150 mM Cl^- . Intracellular pH was unchanged in a non-transfected cell (dashed line; Ctrl) on the same coverslip. **b**, The alkalinization after depolarizing a cell expressing CIC-5 and FaNaC with FMRFamide in the presence of 150 mM Cl^- is reversed when changing to 4 mM Cl^- in the continued presence of FMRFamide. A similar acidification upon low $[Cl^-]_o$ was seen in 11 cells. **c**, A cell expressing CIC-4 and FaNaC was alkalinized by applying FMRFamide. Subsequent Na^+ removal should inhibit CIC-4 by hyperpolarization. Lowering $[Cl^-]_o$ to 4 mM failed to change intracellular pH. Similar results were obtained with eight cells.

respective transport mechanisms probably do not differ as fundamentally as generally assumed when transporters are distinguished from channels. Notably, mutation of the glutamate changed the voltage dependence of CIC-3, CIC-4 and CIC-5 from strongly rectifying to linear^{7,15}. Similar changes in voltage dependence were found for CIC-0 (refs 16, 17) and CIC-K¹⁸ channels, where they were attributed to effects on the gating of a diffusion pore^{1,16}. It is intriguing that the same residue is crucial for coupling Cl^- to H^+ fluxes in CIC-e1, CIC-4 and CIC-5, for the voltage dependence of ion exchange in CIC-3, CIC-4 and CIC-5, and for voltage-dependent gating in CIC-0 (refs 16, 17) and CIC-1 (ref. 19).

Vesicular CLCs are thought to electrically compensate currents of H^+ -ATPases in endosomal/lysosomal compartments, thereby facilitating their acidification^{2,4,15,20–23}. Indeed, the acidification of endosomes and synaptic vesicles was impaired when CIC-5 or CIC-3 were disrupted^{2,20–24}. The highly electrogenic Cl^-/H^+ exchange of CIC-4 and CIC-5 remains compatible with this concept, but the coupling to an H^+ counterflux implies that more metabolic energy is needed for acidification. Unlike Cl^- channels, Cl^-/H^+ exchangers will directly couple Cl^- gradients to vesicular pH gradients. The vesicular Cl^- concentration might influence enzymatic activities²⁵ or might impinge on the osmotic regulation of vesicular volume. Additionally, CIC-4 and CIC-5 might directly acidify endosomes shortly after they pinch off from the plasma membrane by exchanging cytosolic H^+ for luminal Cl^- , which initially is present at the high extracellular concentration. As discussed previously^{1,26}, the role of the steep voltage dependence of CIC-4 and CIC-5 is enigmatic. Our work demonstrates that Cl^-/H^+ exchange activity is not just a peculiarity of a bacterial CLC protein⁸, but rather that a dichotomy between transporters and channels exists within the mammalian CLC family. It additionally suggests a physiological role for vesicular CLCs not only in facilitating endosomal acidification, but also in regulating the Cl^- concentration in endosomal compartments.

METHODS

Perforated patch-clamp and intracellular pH measurements. tsA201 cells (a large T-antigen-expressing derivative of HEK cells) were co-transfected with CLC complementary DNAs (CIC-0, CIC-4, CIC-5, CIC-4 (E224A) and CIC-5 (E211A))

cloned into pCIneo or pCDNA3 expression vectors, and a CD8-encoding plasmid as transfection marker, using Fugene (Roche). Before inserting a dish to the stage of an Olympus BX50WI upright microscope equipped with a $\times 40$ lens, Imago CCD camera and Polychrome IV illumination system (T.I.L.L.), cells were loaded for 15–30 min with BCECF-AM ($1 \mu\text{g ml}^{-1}$; Molecular Probes) in Ringer's solution (145 mM NaCl, 5 mM KCl, 5 mM glucose, 1 mM $MgCl_2$, 1.3 mM Ca-gluconate, 5 mM HEPES, pH 7.4) supplemented with anti-CD8-coated beads (Dyna) at room temperature. Cells were patched in Ringer's solution. When buffered to pH 5 (Fig. 1c), MES replaced HEPES. Patch pipettes connected to the head stage of an Axopatch 200A contained (in mM): 150 KCl, 1 $MgCl_2$, 1.3 $CaCl_2$, 5 HEPES, pH 7.4 and $100 \mu\text{g ml}^{-1}$ gramicidin (Sigma). After gigaseal formation on CD8-positive cells and when gramicidin had lowered the access resistance to <100 – 200 M Ω , the clamp protocol and fluorescence acquisition were started. Owing to the access resistance, the voltages indicated in the corresponding figures are upper estimates of membrane voltages. BCECF was excited at 440 and 480 nm and the ratio of emission at ~ 520 nm was converted to pH after calibration using nigericin and valinomycin ($10 \mu\text{M}$ each) in KCl-based solutions buffered to pH values between 6.5 and 9. Data were analysed using T.I.L.L. Vision, pClamp9 and Origin7.

Intracellular pH measurements on cells depolarized by FaNaC. tsA201 cells were transiently co-transfected with CIC-5 (wild type or E211A mutant)⁷ in pCDNA3 and FaNaC¹⁰ in a bicistronic pIRES vector that co-expresses CD8 (ref. 11). Cells were seeded 24–36 h after transfection on laminin-coated coverslips and measured after 3 h. Cells were loaded with BCECF-AM for ratiometric fluorescence microscopy using an inverted microscope (Zeiss Axiovert 100) with a $\times 100$ oil immersion lens. Fluorescence images were acquired with a CCD camera (Hamamatsu C4742-95) using excitation at 440 nm and 480 nm (Polychrome II, T.I.L.L.). Image acquisition and analysis used Openlab4 (Improvision). FaNaC-transfected cells were identified as above. Cells were continuously perfused with Ringer's solution. Cl^- was replaced by gluconate[−]; Na^+ by NMDG⁺ or by isomolar sucrose when replacing NaCl. FMRFamide was from Bachem. Fluorescence ratios were converted to pH as described above.

Estimation of stoichiometry of Cl^- to H^+ coupling. H^+ fluxes were estimated from cells transfected with CIC-4 or CIC-5 that were clamped to 60 mV as in Fig. 1. $\Delta pH_i/\Delta t$ was converted to the H^+ flux $j(H^+)$ (mols^{−1}) taking the published value of the buffer capacity ($\beta = 47 \pm 2$ mM H^+ per pH unit) of parent HEK cells¹² and estimating the volume of the patched cell from its individual area A , as measured during ion imaging (mean $431 \pm 28 \mu\text{m}^2$), multiplied by the mean height $\bar{h}$ of tsA201 cells ($10.1 \pm 1.0 \mu\text{m}$) measured by confocal stacks according to:

$$j(H^+) = \beta \frac{\Delta pH_i}{\Delta t} A \bar{h}$$

Cl^- fluxes were calculated as $j(Cl^-) = I/F - j(H^+)$, where I is the clamp current and F is Faraday's constant. The apparent coupling ratio $n = j(Cl^-)/j(H^+)$ was then calculated individually for each cell and averaged.

Expression in *Xenopus* oocytes. Capped cRNA was transcribed from CLC constructs cloned in pTLN. Oocytes were prepared, injected and measured as described⁷. Standard extracellular saline was ND96 (105 mM Cl^- , pH 7.4). pH was buffered with HEPES, MES and Tris as appropriate.

Received 11 February; accepted 25 May 2005.

- Jentsch, T. J., Poët, M., Fuhrmann, J. C. & Zdebik, A. A. Physiological functions of CLC Cl^- channels gleaned from human genetic disease and mouse models. *Annu. Rev. Physiol.* **67**, 779–807 (2005).
- Piwon, N., Günther, W., Schwake, M., Bösl, M. R. & Jentsch, T. J. CIC-5 Cl^- channel disruption impairs endocytosis in a mouse model for Dent's disease. *Nature* **408**, 369–373 (2000).
- Lloyd, S. E. *et al.* A common molecular basis for three inherited kidney stone diseases. *Nature* **379**, 445–449 (1996).
- Kornak, U. *et al.* Loss of the CIC-7 chloride channel leads to osteopetrosis in mice and man. *Cell* **104**, 205–215 (2001).
- Kasper, D. *et al.* Loss of the chloride channel CIC-7 leads to lysosomal storage disease and neurodegeneration. *EMBO J.* **24**, 1079–1091 (2005).
- Steinmeyer, K., Schwappach, B., Bens, M., Vandewalle, A. & Jentsch, T. J. Cloning and functional expression of rat CIC-5, a chloride channel related to kidney disease. *J. Biol. Chem.* **270**, 31172–31177 (1995).
- Friedrich, T., Breiderhoff, T. & Jentsch, T. J. Mutational analysis demonstrates that CIC-4 and CIC-5 directly mediate plasma membrane currents. *J. Biol. Chem.* **274**, 896–902 (1999).
- Accardi, A. & Miller, C. Secondary active transport mediated by a prokaryotic homologue of CIC Cl^- channels. *Nature* **427**, 803–807 (2004).
- Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. X-ray structure of a CIC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature* **415**, 287–294 (2002).
- Lingueglia, E., Champigny, G., Lazdunski, M. & Barbry, P. Cloning of the

- amiloride-sensitive FMRamide peptide-gated sodium channel. *Nature* **378**, 730–733 (1995).
11. Poët, M. *et al.* Exploration of the pore structure of a peptide-gated Na⁺ channel. *EMBO J.* **20**, 5595–5602 (2001).
 12. Soleimani, M. *et al.* Pendrin: an apical Cl[−]/OH[−]/HCO₃[−] exchanger in the kidney cortex. *Am. J. Physiol. Renal Physiol.* **280**, F356–F364 (2001).
 13. Roos, A. & Boron, W. F. Intracellular pH. *Physiol. Rev.* **61**, 296–434 (1981).
 14. Myers, V. B. & Haydon, D. A. Ion transfer across lipid membranes in the presence of gramicidin A. II. The ion selectivity. *Biochim. Biophys. Acta* **274**, 313–322 (1972).
 15. Li, X., Wang, T., Zhao, Z. & Weinman, S. A. The ClC-3 chloride channel promotes acidification of lysosomes in CHO-K1 and Huh-7 cells. *Am. J. Physiol. Cell Physiol.* **282**, C1483–C1491 (2002).
 16. Dutzler, R., Campbell, E. B. & MacKinnon, R. Gating the selectivity filter in ClC chloride channels. *Science* **300**, 108–112 (2003).
 17. Traverso, S., Elia, L. & Pusch, M. Gating competence of constitutively open CLC-0 mutants revealed by the interaction with a small organic inhibitor. *J. Gen. Physiol.* **122**, 295–306 (2003).
 18. Waldegger, S. & Jentsch, T. J. Functional and structural analysis of ClC-K chloride channels involved in renal disease. *J. Biol. Chem.* **275**, 24527–24533 (2000).
 19. Fahlke, C., Yu, H. T., Beck, C. L., Rhodes, T. H. & George, A. L. Jr Pore-forming segments in voltage-gated chloride channels. *Nature* **390**, 529–532 (1997).
 20. Günther, W., Piwon, N. & Jentsch, T. J. The ClC-5 chloride channel knock-out mouse — an animal model for Dent's disease. *Pflügers Arch.* **445**, 456–462 (2003).
 21. Stobrawa, S. M. *et al.* Disruption of ClC-3, a chloride channel expressed on synaptic vesicles, leads to a loss of the hippocampus. *Neuron* **29**, 185–196 (2001).
 22. Hara-Chikuma, M. *et al.* ClC-3 chloride channels facilitate endosomal acidification and chloride accumulation. *J. Biol. Chem.* **280**, 1241–1247 (2005).
 23. Yoshikawa, M. *et al.* CLC-3 deficiency leads to phenotypes similar to human neuronal ceroid lipofuscinosis. *Genes Cells* **7**, 597–605 (2002).
 24. Hara-Chikuma, M., Wang, Y., Guggino, S. E., Guggino, W. B. & Verkman, A. S. Impaired acidification in early endosomes of ClC-5 deficient proximal tubule. *Biochem. Biophys. Res. Commun.* **329**, 941–946 (2005).
 25. Davis-Kaplan, S. R., Askwith, C. C., Bengtzen, A. C., Radisky, D. & Kaplan, J. Chloride is an allosteric effector of copper assembly for the yeast multicopper oxidase Fet3p: an unexpected role for intracellular chloride channels. *Proc. Natl Acad. Sci. USA* **95**, 13641–13645 (1998).
 26. Jentsch, T. J., Stein, V., Weinreich, F. & Zdebek, A. A. Molecular structure and physiological function of chloride channels. *Physiol. Rev.* **82**, 503–568 (2002).

Acknowledgements We thank M. Lazdunski for the gift of the FaNaC-CD8 expression vector, and M. Petersen and P. Breiden for technical assistance. This work was supported in part by the Prix Louis-Jeantet de Médecine.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.J.J. (Jentsch@zmnh.uni-hamburg.de).

LETTERS

SUMO-modified PCNA recruits Srs2 to prevent recombination during S phase

Boris Pfander¹, George-Lucian Moldovan¹, Meik Sacher¹, Carsten Hoege^{1†} & Stefan Jentsch¹

Damaged DNA, if not repaired before replication, can lead to replication fork stalling and genomic instability^{1–3}; however, cells can switch to different damage bypass modes that permit replication across lesions. Two main bypasses are controlled by ubiquitin modification of proliferating cell nuclear antigen (PCNA), a homotrimeric DNA-encircling protein that functions as a polymerase processivity factor and regulator of replication-linked functions^{4,5}. Upon DNA damage, PCNA is modified at the conserved lysine residue 164 by either mono-ubiquitin or a lysine-63-linked multi-ubiquitin chain⁵, which induce error-prone or error-free replication bypasses of the lesions^{5,6}. In S phase, even in the absence of exogenous DNA damage, yeast PCNA can be alternatively modified by the small ubiquitin-related modifier protein SUMO⁵; however the consequences of this remain controversial^{5–7}. Here we show by genetic analysis that SUMO-modified PCNA functionally cooperates with Srs2, a helicase that blocks recombinational repair by disrupting Rad51 nucleoprotein filaments^{8,9}. Moreover, Srs2 displays a preference for interacting directly with the SUMO-modified form of PCNA, owing to a specific binding site in its carboxy-terminal tail. Our finding suggests a model in which SUMO-modified PCNA recruits Srs2 in S phase in order to prevent unwanted recombination events of replicating chromosomes.

Modification of PCNA by ubiquitin involves enzymes of the RAD6 DNA damage bypass mode (Fig. 1a; see also ref. 5). Whereas mono-ubiquitination of PCNA requires Rad6 and Rad18, modification by lysine (K)-63-linked multi-ubiquitin chains additionally requires the heterodimer Ubc13–Mms2 and Rad5. Yeast PCNA (Pol30) is also modified by SUMO at K164, and, to a lesser extent, at the non-conserved residue K127 (ref. 5), which matches the ΨKxD/E consensus motif for SUMO modification (SUMOylation)¹⁰ and Ubc9 binding¹¹ (Ψ represents an aliphatic amino acid). SUMOylation depends on Ubc9, but modification at K164 only requires the SUMO ligase Siz1 (ref. 12) (Fig. 1b). All of these PCNA modifications are limited to S phase (ref. 5; see Supplementary Fig. 1); in contrast to ubiquitination, SUMOylation occurs even in the absence of exogenous damage⁵.

The role of PCNA SUMOylation has remained elusive. We noticed previously⁵ that in *Saccharomyces cerevisiae*, SUMOylation of PCNA at K127 and K164 is detrimental to DNA damage tolerance in the absence of PCNA ubiquitination. However, it has been suggested by genetic arguments⁶ that, analogous to the model^{5,6,13,14} for mono-ubiquitinated PCNA, SUMOylation promotes error-prone synthesis through recruitment of a translesion polymerase. To address this issue we used the *siz1* mutant, which abolishes PCNA SUMOylation at the K164 site but leaves PCNA ubiquitination unaffected (Fig. 1b). As noted previously⁶, this mutant partially suppresses the hypersensitivity to ultraviolet light or the DNA-alkylating drug methylmethane sulphonate (MMS) of *rad6*, *rad18*, *rad5* and *mms2* mutants

(Fig. 1c, d; see also Supplementary Fig. 2a–c). Notably, this effect of *siz1* is linked to a deficiency in PCNA SUMOylation, because hypersensitivity suppression of *rad6* and *rad18* by a PCNA mutant that lacks K164 (*pol30-K164R*) was identical to suppression by *siz1* (Fig. 1e, f). A PCNA mutant lacking both SUMOylation sites (*pol30-RR*) suppressed *rad6* and *rad18* to an even greater extent, indicating that SUMOylation at K127 and K164 additively inhibits a salvage pathway.

We identified RAD52—an essential upstream element of recombinational repair^{15,16}—as a high-dose suppressor of the MMS hypersensitivity of the PCNA mutant *pol30-K164R* (data not shown). Thus, cells deficient in PCNA modification can survive DNA damage by activating a RAD52-dependent recombinational bypass. Conversely, suppression of the ultraviolet sensitivity of *rad6* and *rad18* mutants by the *siz1* mutation or PCNA lysine mutations strictly requires a functional RAD52 pathway (Fig. 1c–f). In fact, hypersensitivity suppression was not only absent in *rad52* mutants, but also in *rad51*, *rad54* and *rad55* mutants (Supplementary Fig. 2d), which are defective in the central activity of recombination^{15,16}. In contrast, Rad50—which is required before strand exchange in DNA strand resection—and Rad59—a Rad52 homologue—are apparently not needed for this salvage pathway (Supplementary Fig. 2e). Moreover, hypersensitivity suppression of *rad6* by *siz1* was not affected by eliminating the RAD2 nucleotide excision repair pathway (Supplementary Fig. 2f). From these findings we conclude that SUMOylation of PCNA blocks specifically a recombination pathway and that SUMOylation of both K164 and K127 contribute to the inhibition.

Our conclusion that PCNA SUMOylation blocks a RAD52-dependent pathway is in line with a recent model⁷. This model was based on the remark that *rad6 rad52* and *pol30-K164R rad52* have similar phenotypes⁷ (instead of comparing *rad6 rad52* with the triple mutant *rad6 pol30-K164R rad52*). However, this is not the case, as *rad6 rad52* is in fact more sensitive than *pol30-K164R rad52* (Supplementary Fig. 2g), which reflects the fact that Rad6 has other substrates in addition to PCNA (refs 5, 7; Fig. 1e; see also Supplementary Fig. 2h). In contrast, PCNA is phenotypically the essential target for the Rad18, Rad5 and Mms2 enzymes⁵. We also noticed that SUMOylation of PCNA seems to have another function besides its main role in preventing a recombinational repair pathway. This currently undefined activity is primarily perceptible in double mutants of *rad18* with *rad52* (or *rad51*, *rad54*, *rad55*), which are rendered more sensitive to ultraviolet light by the absence of Siz1 or by PCNA lysine mutants (Fig. 1d, f and Supplementary Fig. 2d; see also Fig. 3 and Supplementary Fig. 4 for this effect in *siz1 srs2* double mutants).

We realized that the inhibitory function of PCNA SUMOylation on a RAD52 pathway is similar to the known role of the helicase Srs2 (also known as Hpr5)^{17–22}. This enzyme is a potent inhibitor of

¹Department of Molecular Cell Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. †Present address: Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany.

recombination as it disrupts Rad51 nucleoprotein filaments^{8,9}, which are crucial early recombinogenic intermediates. Moreover, it has been noted²³ that *srs2Δ* is lethal in combination with a de-SUMOylation enzyme mutant. To explore potential links between PCNA SUMOylation and Srs2 we looked for a physical interaction. In two-hybrid assays, full-length Srs2 is inactive due to low expression, but an amino-terminally truncated Srs2 fragment lacking the helicase domain (Srs2ΔN) is able to bind Rad51 (ref. 9; Fig. 2a). Srs2ΔN also bound Rad18 and Rad5, and, importantly, PCNA as well. Moreover, Srs2ΔN bound SUMO²⁴ (Fig. 2a), and showed greater affinity for binding the wild-type form of SUMO (Smt3GG) that can form conjugates via its C-terminal di-glycine motif. Notably, the PCNA fusion used for the two-hybrid assays (BD-PCNA) was modified by SUMO at both lysine residues, and SUMOylation at K164 depended on Siz1 (Fig. 2b). When we introduced the lysine mutations in the BD-PCNA construct (as single mutations or in combination) we observed a stepwise reduction in Srs2ΔN binding in two-hybrid assays that was proportional to defects in SUMOylation (Fig. 2c). Furthermore, two-hybrid interaction between the two proteins was also reduced in *Siz1*-deficient cells. From these data we conclude that Srs2 binds PCNA *in vivo*, and that the interaction is strongly augmented by PCNA SUMOylation.

We confirmed these interactions by pull-down assays using a glutathione S-transferase (GST) fusion of Srs2ΔN (Fig. 2d). Because of the activity of de-SUMOylating enzymes in the cell extract we could not observe the interaction between Srs2 and the modified form of PCNA unless we triggered PCNA SUMOylation by addition

of 0.3% MMS⁵ to the medium. Under these conditions GST-Srs2ΔN pulled down the SUMOylated forms of PCNA with greater preference compared with the unmodified form (Fig. 2e, f). In fact, Srs2 interacted strongly with PCNA modified by SUMO at K127 or K164, and especially with the doubly modified form. An excess of free SUMO, but not ubiquitin, provided competition for binding of SUMOylated PCNA to Srs2 (Fig. 2g), emphasizing that it is specifically SUMO that is relevant for binding. Furthermore, we SUMOylated purified recombinant PCNA *in vitro* at both K164 (which depends on Siz1) and K127, and found also in this set-up that GST-Srs2ΔN also pulled down the SUMOylated forms of PCNA with preference (Fig. 2h), suggesting that Srs2 binds SUMOylated PCNA directly. Notably, the steady-state level of SUMOylated PCNA was induced *in vivo* by overexpression of Srs2ΔN, suggesting that the PCNA-SUMO conjugate is shielded by Srs2ΔN against the activity of de-SUMOylating enzymes (Supplementary Fig. 3). We mapped the interaction with PCNA and SUMO to the C-terminal 138 residues of Srs2's tail (Fig. 2i); C-terminal truncations by as few as six amino acids strongly reduced the two-hybrid interaction with both proteins (Fig. 2i). Notably, this region of Srs2 is distinct from the Rad51 interaction domain, which is located further amino-terminally on Srs2's tail (Fig. 2i).

Given the physical interaction between Srs2 and SUMOylated PCNA we examined whether the proteins functionally cooperate. When we assayed for the suppression of the ultraviolet and MMS sensitivities of *rad18* or *rad6* mutants by *srs2*, or by mutants defective in PCNA SUMOylation, we noticed that *rad18 srs2*, *rad18 siz1*,

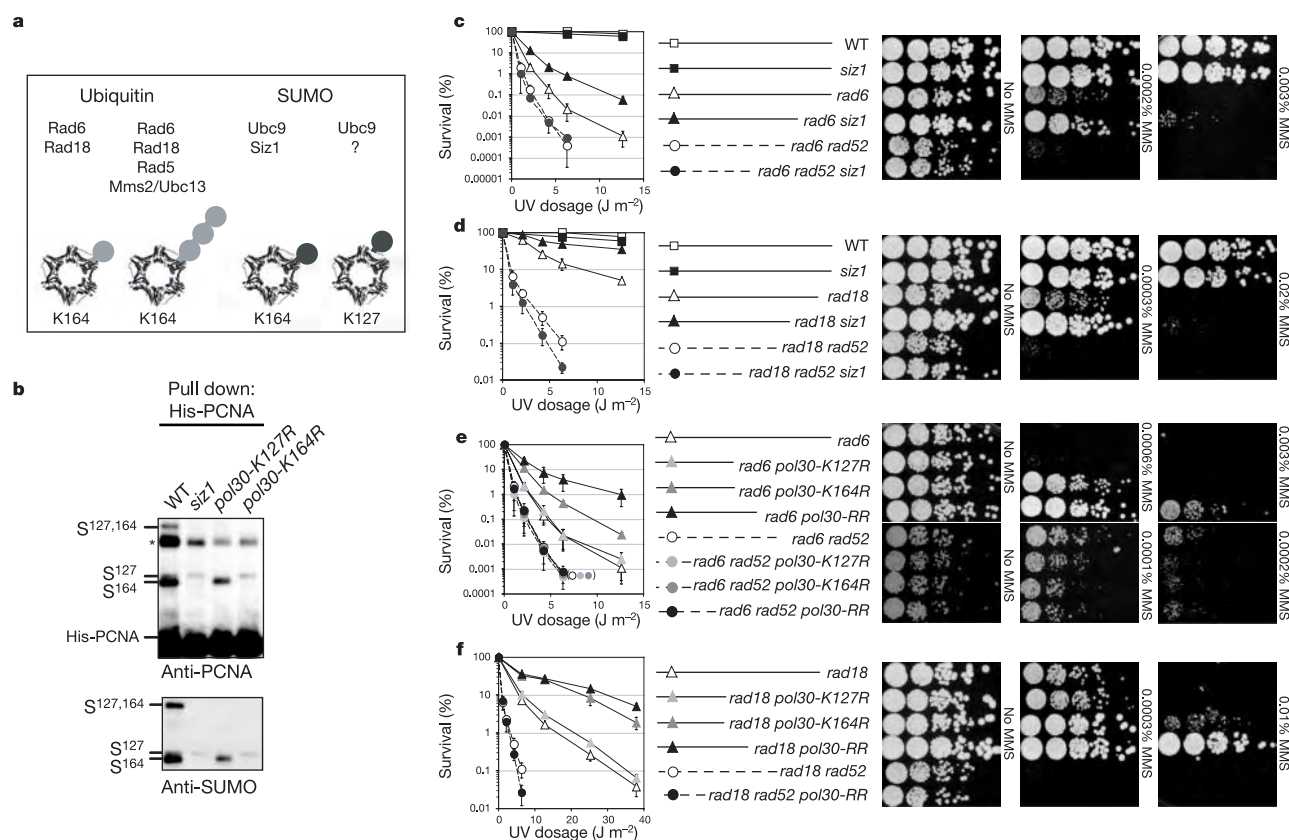


Figure 1 | Suppression of DNA damage sensitivity of RAD6 pathway mutants by deficiency in PCNA SUMOylation requires homologous recombination. **a**, PCNA modifications in *S. cerevisiae* and the required enzymes. **b**, Siz1 mediates PCNA SUMOylation at K164 (S¹⁶⁴), but not at K127. PCNA and modified forms were enriched by His-PCNA pull down. The asterisk denotes a cross-reacting band. **c–f**, Suppression of *rad6* or *rad18*

DNA damage sensitivity by *siz1* or PCNA lysine mutants is dependent on RAD52. Shown are serial dilutions of cells on plates containing MMS as indicated (right), and the quantification of survival rates after ultraviolet irradiation (left). Error bars indicate standard deviations of independent experiments.

rad18 srs2 siz1 and *rad18 srs2 pol30-RR* mutants all showed nearly identical phenotypes (Fig. 3b). This epistatic relationship reveals that PCNA SUMOylation and Srs2 function in the same pathway. Importantly, *srs2ΔC* mutant alleles (*srs2ΔC136*, *srs2ΔC24*, *srs2ΔC6*)

that express truncated proteins defective in binding to SUMOylated PCNA suppressed the damage sensitivity of *rad18* equally and epistatically to the *siz1* mutant defective in PCNA SUMOylation (Fig. 3d; see also Supplementary Fig. 4e). The suppression of

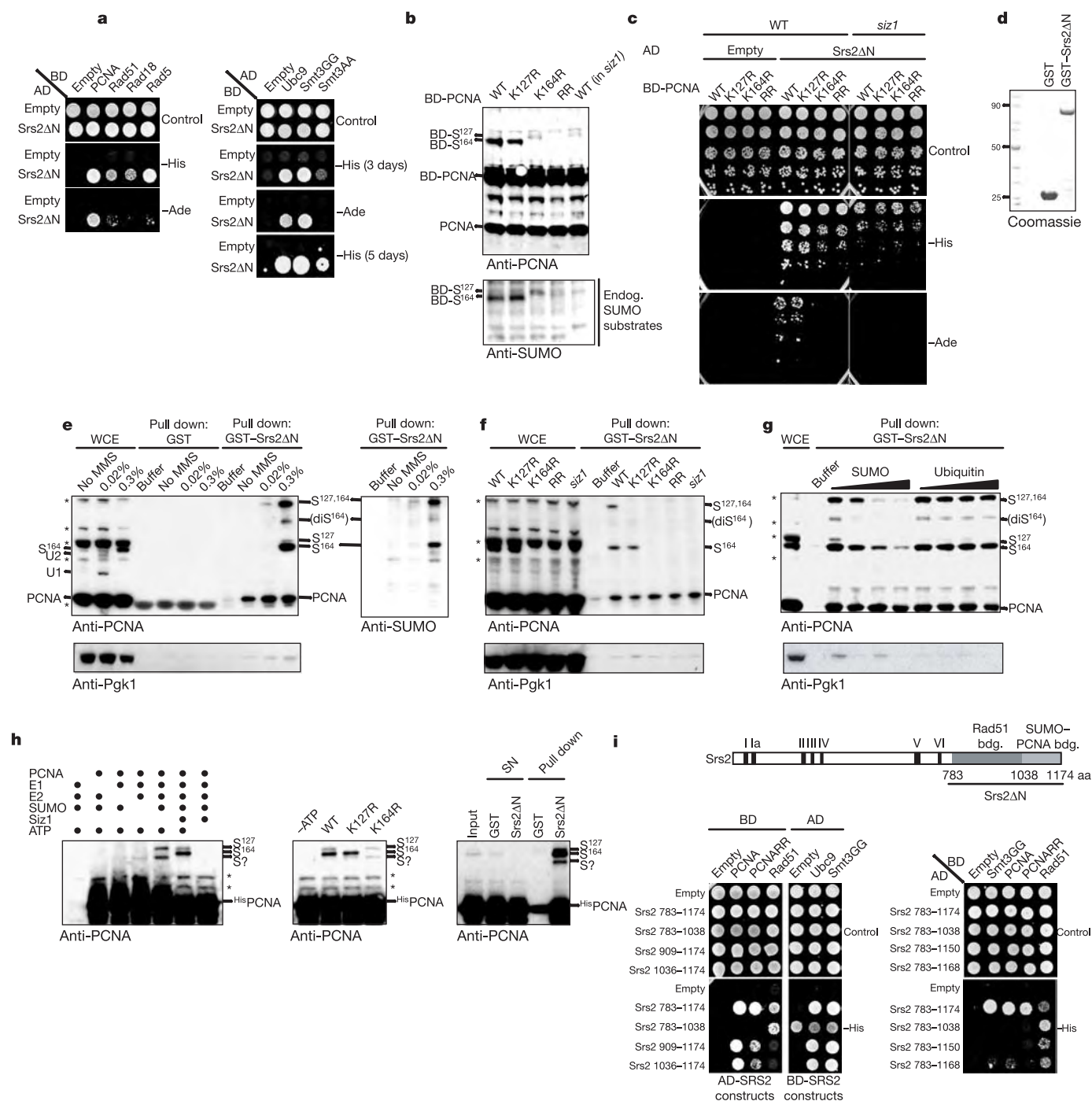


Figure 2 | Srs2 preferentially binds SUMOylated PCNA. **a**, Two-hybrid interactions of a C-terminal fragment of Srs2 (Srs2ΔN) with PCNA, Rad51, Rad18, Rad5 (left), Ubc9 and Smt3 (SUMO) (right) were identified on selective media (–His, –Ade). Fusions with activating domain (AD) or DNA-binding domain (BD) are indicated. **b**, BD-PCNA is SUMOylated on K164 (BD-S¹⁶⁴; Siz1-dependent) and K127 (BD-S¹²⁷) in the presence of AD-Srs2ΔN. **c**, Two-hybrid interaction of PCNA mutant proteins with Srs2ΔN in wild-type or *siz1* cells. **d–g**, Pull-down experiments with GST-Srs2ΔN or GST from extracts of wild-type cells treated with different concentrations of MMS (**e, g**), PCNA lysine mutants or *siz1* cells (**f**). **g**, Binding was competed by SUMO (Smt3), but not ubiquitin. Bands in **e–g**

correspond to unmodified (PCNA), ubiquitinated (U1, U2) and SUMO-modified (S) PCNA, and PCNA carrying a putative SUMO–SUMO chain at K164 (diS¹⁶⁴). Asterisks denote cross-reacting bands. **h**, PCNA was SUMOylated *in vitro*. In addition to K127 and K164, SUMOylation *in vitro* also occurs at another site (S?). Reaction products were subjected to pull downs using GST-Srs2ΔN. An excess of Ubc9 (which also binds Srs2) did not reinforce binding of Srs2 to SUMOylated PCNA (not shown). **i**, Mapping of the binding sites of PCNA, Rad51, Ubc9 and SUMO using truncated Srs2 constructs by two-hybrid analysis. Regions I–VI are helix homology boxes¹⁸.

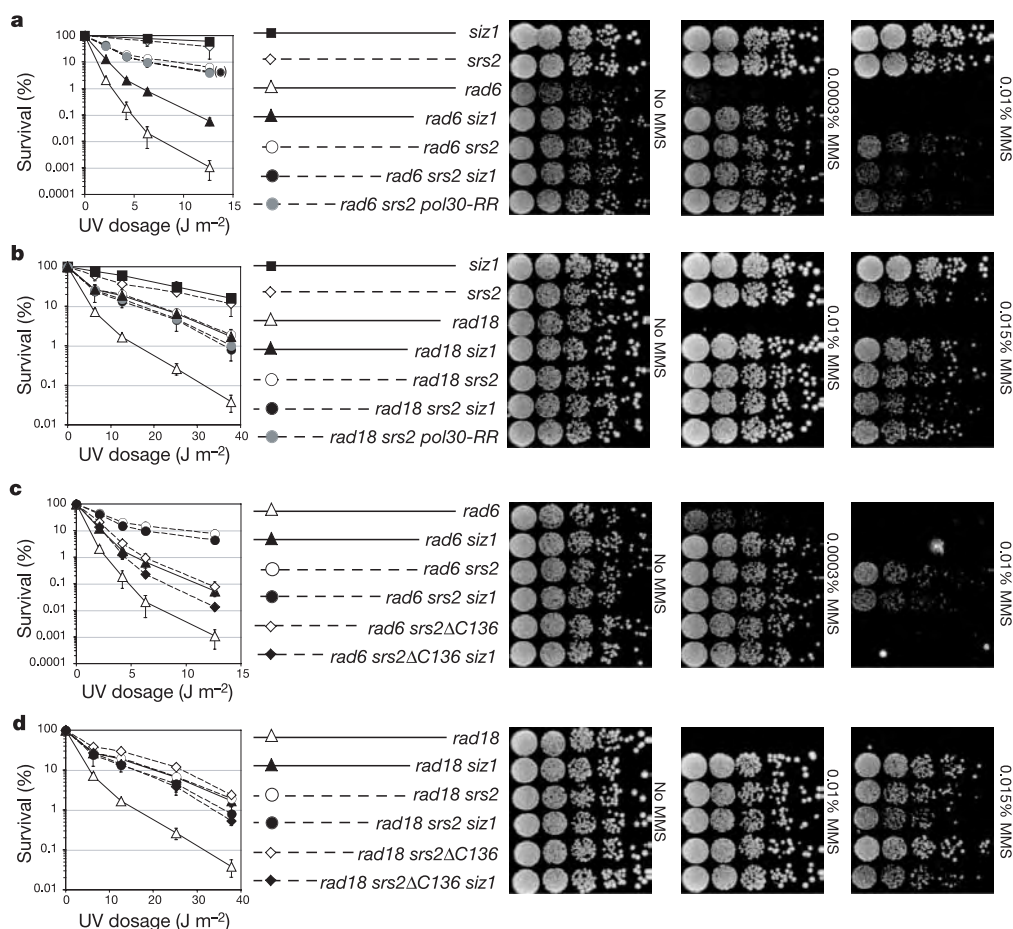


Figure 3 | Mutants deficient in PCNA SUMOylation and in the C-terminal tail of Srs2 (*srs2ΔC*) are epistatic with respect to suppression of *RAD6* pathway mutants. a–d, Survival rates after ultraviolet light irradiation and sensitivity to MMS were determined as in Fig. 1c–f. Error bars represent standard deviations of independent experiments. The C-terminal

truncations of Srs2 are expressed at wild-type levels. The nuclear localization signal (amino acids 1117–1121; ref. 28) is not essential for Srs2's nuclear functions, as *srs2Δ136* and *srs2Δ6* behave identically in all genetic assays, but are different compared with *srs2Δ* (for example, in the *rad6* background; see also Fig. 4 and Supplementary Figs 4d, e and 5).

ultraviolet and MMS sensitivities for the *rad6* mutant by *srs2* was greater than by mutants deficient in PCNA SUMOylation (Fig. 3a), indicating that *srs2* suppresses *rad6* by two mechanisms: a PCNA–SUMO-dependent and -independent mode. Notably, in contrast to the *srs2* knockout, *srs2ΔC* mutants rescued *rad6* only partially and precisely to the level of suppression by *siz1* (Fig. 3c; see also Supplementary Fig. 4d). This finding underscores the importance of Srs2's C-terminal tail in binding to SUMOylated PCNA, and also illustrates that other functions of Srs2 are unaffected by the *srs2ΔC* mutations.

The PCNA–SUMO–Srs2 check is not restricted to haploids, because *siz1* and *srs2ΔC* mutants also suppressed the sensitivity of diploid *rad18* mutants (Supplementary Fig. 5a, b). Moreover, these mutants are, in contrast to the *srs2Δ* mutant, only mildly sensitive to DNA-damaging agents (Supplementary Fig. 5a; see also ref. 18). This indicates that recombination between homologous chromosomes, in which Srs2 is involved, is not influenced by this distinctive SUMO-dependent Srs2 pathway. Similarly, in contrast to the *srs2Δ* diploid, neither the corresponding *srs2ΔC* mutants nor the mutants defective in PCNA SUMOylation displayed hyperactive interchromosomal recombination (Fig. 4a; see also ref. 25). In fact, these mutants suppressed the high interchromosomal recombination rate of the *rad18* mutant (Fig. 4a), which probably arises through its inability to replicate across lesions. This suggests that the *RAD52* salvage pathway that is unleashed by the absence of the

PCNA–SUMO–Srs2 check is probably not interchromosomal recombination.

As an alternative, we speculated that the identified salvage pathway might involve recombination between sister chromatids during S phase. As intrachromosomal recombination between direct repeats is to some extent caused by sister chromatid recombination¹⁵, we assayed for this activity using a set-up that can differentiate between gene conversion and recombination-mediated deletions²⁰. Indeed, we observed that the *siz1* mutant defective in PCNA SUMOylation and *srs2ΔC* had increased recombination rates, which were further increased by a deficiency in *RAD18* (Fig. 4b). Thus, the recombination pathway that is set free in the absence of the PCNA–SUMO–Srs2 pathway has the characteristics of sister chromatid recombination. Intrachromosomal recombination was particularly upregulated by *srs2Δ*, however largely through gene conversion (ref. 26 and Fig. 4b). This probably reflects another known function of Srs2, namely to regulate the length of conversion tracts during recombination²⁶, which is independent of PCNA SUMOylation.

RAD6 pathway mutants exhibit a mutator phenotype, and because this phenotype of *rad18* is suppressed by *siz1*, it has been suggested⁶ that SUMOylated PCNA might recruit translesion polymerases. However, as this effect of SUMO is only present in a mutant background in which PCNA cannot be ubiquitinated (wild-type and *siz1* cells show identical spontaneous mutagenesis rates), we hypothesized that mutations might arise through faulty replication

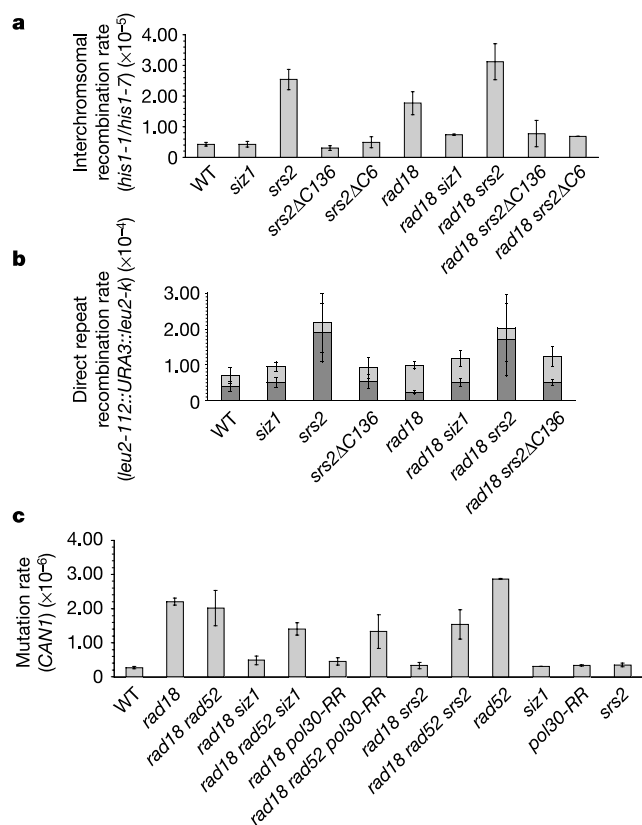


Figure 4 | Influence of the PCNA-SUMO-Srs2 check on recombination and mutator phenotypes. **a**, The C-terminal tail of Srs2 is not responsible for the increased interchromosomal recombination rate (measured in *his1-1/his1-7* diploids) of *srs2Δ*, but is, together with PCNA SUMOylation, required for the phenotype of *rad18*. **b**, PCNA SUMOylation inhibits the intrachromosomal recombination between direct repeats (*leu2-112* and *leu2-k* alleles, flanking *URA3*). Overall recombination (*LEU*⁺) and gene conversion rates (*LEU*⁺*URA*⁺, darker grey sections) are shown. **c**, The mutator phenotype of *rad18* depends on PCNA SUMOylation and the C-terminal tail of Srs2. Rates of forward mutation of the *CAN1* locus were determined. In **a–c**, the mean values of two to six independent fluctuation tests are shown. Error bars indicate standard deviations.

if the *RAD6* bypass is inactivated by mutation and the *RAD52*-dependent salvage pathway is additionally blocked by PCNA SUMOylation. In fact, the mutator phenotype of *rad18* was not only suppressed by *siz1* but also by *pol30-RR* and *srs2* mutants (Fig. 4c). Additional deletion of *RAD52* strongly neutralized the suppression, indicating that inhibition of the *RAD52* salvage pathway is the major cause of the mutator phenotype of *rad18*.

This and our previous work⁵ emphasize the importance of PCNA modifications for decision-making at the replication fork. Whereas PCNA ubiquitination mediates post-replicative lesion bypasses^{5,6,13,14}, modification with SUMO, which occurs even in the absence of exogenous DNA damage, seems to be a guarding mechanism that prevents unwanted recombination during replication. SUMOylation and ubiquitination might represent autonomous triggers, which operate independently from each other. However, as components of the two modification pathways interact (ref. 5 and Fig. 2a), they might build a switchboard in which the PCNA-SUMO-Srs2 check facilitates channelling into the *RAD6*-dependent bypass.

METHODS

Protein techniques. Yeast native extract was prepared by glass bead lysis in 150 mM NaCl, 50 mM Tris-HCl pH 7.4 supplemented with protease inhibitors,

followed by detergent extraction (1% Triton X-100, 0.05% SDS) and pre-clearing by centrifugation. For the pull-down assay, 50 μg of GST or GST-Srs2ΔN bound to beads were incubated with 2.5 mg of yeast native lysate for 3 h at 4 °C. Beads were then washed and eluted in sample buffer containing 8 M urea. Whole-cell extract samples correspond to 1/50 of the total input. For competition experiments, recombinant free SUMO (Smt3) or ubiquitin was added to the beads together with the lysate to final concentrations of 0, 4, 40 and 400 μM, respectively. Yeast protein extracts and analytical denaturing NiNTA pull down were done as described⁵.

Sensitivity, mutagenesis and recombination assays. For qualitative analysis of MMS sensitivity, cells from overnight cultures were spotted on YPD plates containing MMS. For quantification of ultraviolet sensitivities, fixed amounts of cells were irradiated with different dosages of ultraviolet light (254 nm) after plating on YPD plates. Colony-forming units were counted after 3 days of growth in the dark. Values are averages from three to seven independent experiments using duplicates. Spontaneous forward mutation of the *CAN1* locus was measured by growth on complete media lacking arginine with 60 mg l⁻¹ canavanine⁶. Interchromosomal recombination between the hetero-alleles *his1-1* and *his1-7* in diploid cells was determined by growth on complete media without histidine. For direct repeat recombination a strain derived from 344-109D (provided by H. L. Klein) was used that contains a *leu2-112::URA3::leu2-k* array in W303 *RAD5*⁺ background. This was used to score for overall recombination (growth on medium lacking leucine) and gene conversion (growth on medium lacking leucine and uracil). Mutational and recombinational rates were determined by fluctuation analysis²⁷. Briefly, ten independent cultures were grown to saturation then plated on selective and non-selective plates. The number of mutational/recombinational events was determined using a maximum-likelihood approach for the deviation of the number of mutants. Averages were obtained from at least two independent experiments. For interchromosomal recombination rates the spontaneous reversion rates of the respective homozygous *his1-1/his1-1* and *his1-7/his1-7* strains were subtracted.

Additional methods, materials and strains are presented in Supplementary Information.

Received 29 November 2004; accepted 15 April 2005.

Published online 1 June 2005.

- Barbour, L. & Xiao, W. Regulation of alternative replication bypass pathways at stalled replication forks and its effects on genome stability: a yeast model. *Mutat. Res.* **532**, 137–155 (2003).
- Smirnova, M. & Klein, H. L. Role of the error-free damage bypass postreplication repair pathway in the maintenance of genomic stability. *Mutat. Res.* **532**, 117–135 (2003).
- Osborn, A. J., Elledge, S. J. & Zou, L. Checking on the fork: the DNA-replication stress-response pathway. *Trends Cell Biol.* **12**, 509–516 (2002).
- Tsurimoto, T. PCNA binding proteins. *Front. Biosci.* **4**, D849–D858 (1999).
- Hoegel, C., Pfander, B., Moldovan, G. L., Pyrowolakis, G. & Jentsch, S. RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. *Nature* **419**, 135–141 (2002).
- Stelter, P. & Ulrich, H. D. Control of spontaneous and damage-induced mutagenesis by SUMO and ubiquitin conjugation. *Nature* **425**, 188–191 (2003).
- Haracska, L., Torres-Ramos, C. A., Johnson, R. E., Prakash, S. & Prakash, L. Opposing effects of ubiquitin conjugation and SUMO modification of PCNA on replicational bypass of DNA lesions in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **24**, 4267–4274 (2004).
- Veaute, X. et al. The Srs2 helicase prevents recombination by disrupting Rad51 nucleoprotein filaments. *Nature* **423**, 309–312 (2003).
- Krejci, L. et al. DNA helicase Srs2 disrupts the Rad51 presynaptic filament. *Nature* **423**, 305–309 (2003).
- Johnson, E. S. & Blobel, G. Cell cycle-regulated attachment of the ubiquitin-related protein SUMO to the yeast septins. *J. Cell Biol.* **147**, 981–994 (1999).
- Bernier-Villamor, V., Sampson, D. A., Matunis, M. J. & Lima, C. D. Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. *Cell* **108**, 345–356 (2002).
- Johnson, E. S. & Gupta, A. A. An E3-like factor that promotes SUMO conjugation to the yeast septins. *Cell* **106**, 735–744 (2001).
- Kannouche, P. L., Wing, J. & Lehmann, A. R. Interaction of human DNA polymerase η with monoubiquitinated PCNA: a possible mechanism for the polymerase switch in response to DNA damage. *Mol. Cell* **14**, 491–500 (2004).
- Watanabe, K. et al. Rad18 guides poleta to replication stalling sites through physical interaction and PCNA monoubiquitination. *EMBO J.* **23**, 3886–3896 (2004).
- Symington, L. S. Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiol. Mol. Biol. Rev.* **66**, 630–670 (2002).

16. Aylon, Y. & Kupiec, M. New insights into the mechanism of homologous recombination in yeast. *Mutat. Res.* **566**, 231–248 (2004).
17. Lawrence, C. W. & Christensen, R. B. Metabolic suppressors of trimethoprim and ultraviolet light sensitivities of *Saccharomyces cerevisiae* rad6 mutants. *J. Bacteriol.* **139**, 866–887 (1979).
18. Aboussekhra, A. et al. RADH, a gene of *Saccharomyces cerevisiae* encoding a putative DNA helicase involved in DNA repair. Characteristics of radH mutants and sequence of the gene. *Nucleic Acids Res.* **17**, 7211–7219 (1989).
19. Schiestl, R. H., Prakash, S. & Prakash, L. The SRS2 suppressor of rad6 mutations of *Saccharomyces cerevisiae* acts by channeling DNA lesions into the RAD52 DNA repair pathway. *Genetics* **124**, 817–831 (1990).
20. Palladino, F. & Klein, H. L. Analysis of mitotic and meiotic defects in *Saccharomyces cerevisiae* SRS2 DNA helicase mutants. *Genetics* **132**, 23–37 (1992).
21. Ulrich, H. D. The srs2 suppressor of UV sensitivity acts specifically on the RAD5- and MMS2-dependent branch of the RAD6 pathway. *Nucleic Acids Res.* **29**, 3487–3494 (2001).
22. Broomfield, S. & Xiao, W. Suppression of genetic defects within the RAD6 pathway by srs2 is specific for error-free post-replication repair but not for damage-induced mutagenesis. *Nucleic Acids Res.* **30**, 732–739 (2002).
23. Soustelle, C. et al. A new *Saccharomyces cerevisiae* strain with a mutant Smt3-deconjugating Ulp1 protein is affected in DNA replication and requires Srs2 and homologous recombination for its viability. *Mol. Cell. Biol.* **24**, 5130–5143 (2004).
24. Uetz, P. et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
25. Friedl, A. A., Liefshitz, B., Steinlauf, R. & Kupiec, M. Deletion of the SRS2 gene suppresses elevated recombination and DNA damage sensitivity in rad5 and rad18 mutants of *Saccharomyces cerevisiae*. *Mutat. Res.* **486**, 137–146 (2001).
26. Aguilera, A. & Klein, H. L. Genetic control of intrachromosomal recombination in *Saccharomyces cerevisiae*. I. Isolation and genetic characterization of hyper-recombination mutations. *Genetics* **119**, 779–790 (1988).
27. Rosche, W. A. & Foster, P. L. Determining mutation rates in bacterial populations. *Methods* **20**, 4–17 (2000).
28. Rong, L. & Klein, H. L. Purification and characterization of the SRS2 DNA helicase of the yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.* **268**, 1252–1259 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank U. Cramer for technical assistance, S. Kumar for the gift of the SRS2ΔN clone, D. Siepe for computational analysis, and E. S. Johnson, H. L. Klein, C. Pohl, H. Richly and H. D. Ulrich for materials. This work is supported (to S. J.) by the Max Planck Society, Deutsche Krebshilfe, Deutsche Forschungsgemeinschaft, and Fonds der chemischen Industrie.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.J. (Jentsch@biochem.mpg.de).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

A tangential route to success

Sometimes tangents can lead to promising new science, as well as fresh career opportunities. That's the lesson learned by Renos Savva, this year's winner of London First's Young Biotechnology Entrepreneur Award, announced last week.

When Savva was doing his PhD at University College London, he was mostly concerned with how cells repair their DNA after it becomes damaged. But when he joined a research group around the corner at Birkbeck College in 1999, his colleagues directed his attention towards structural biology — a step away from the linear world of genetics and genomics into the three-dimensional domain of proteomics.

At the time, the group was developing technology to clone and express proteins. Traditional methods emphasized bioinformatics to guess how differences in sequences could result in different protein shapes and folds. But Savva and his colleagues felt that the approach was incomplete, because it was only educated guesswork. So they found a way to randomly sample the genome — using damaged DNA, Savva's speciality —

and then clone a library of proteins that would better represent these changes.

The work had only "tenuous links" to Savva's PhD work, he says, but he nevertheless was prepared to take the plunge and follow in the footsteps of his father, also an entrepreneur. After the university helped the group obtain patent protection, the researchers spun out a company, struggled for seed money, then searched for clients. Now Savva splits his time between being research director of the company Domainex and as a research director at Birkbeck.

Savva's experience shows that left turns can lead to profitable places. But he says that tangents alone are not enough. Scientists who want to enter business also need to be comfortable with risks, and to have infrastructure in place to turn academic ideas into business opportunities.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

US Head Office, New York
 345 Park Avenue South, 10th Floor,
 New York, NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager/Corporations:
 Peter Bless
 Classified Sales Representatives
 Tel: +1 800 989 7718

**New York/Pennsylvania/
 Latin America:** Kelly Roman
**Midwest USA/Maryland/
 NIH:** Wade Tucker
East USA/Canada:
 Janine Taormina

San Francisco Office
Classified Sales Representative:
 Michaela Bjorkman
 225 Bush Street, Suite 1453
 San Francisco, CA 94104
 Tel: +1 415 781 3803
 Fax: +1 415 781 3805
 e-mail: m.bjorkman@naturesf.com

European Head Office, London
 The Macmillan Building,
 4 Crinan Street,
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director: Nevin Bayoumi (4978)
European Sales Manager: Andy Douglas (4975)

Advertising Production Manager: Billie Franklin
 To send materials use London address above.
 Tel: +44 (0) 20 7843 4814
 Fax: +44 (0) 20 7843 4814
 e-mail: naturejobs@nature.com

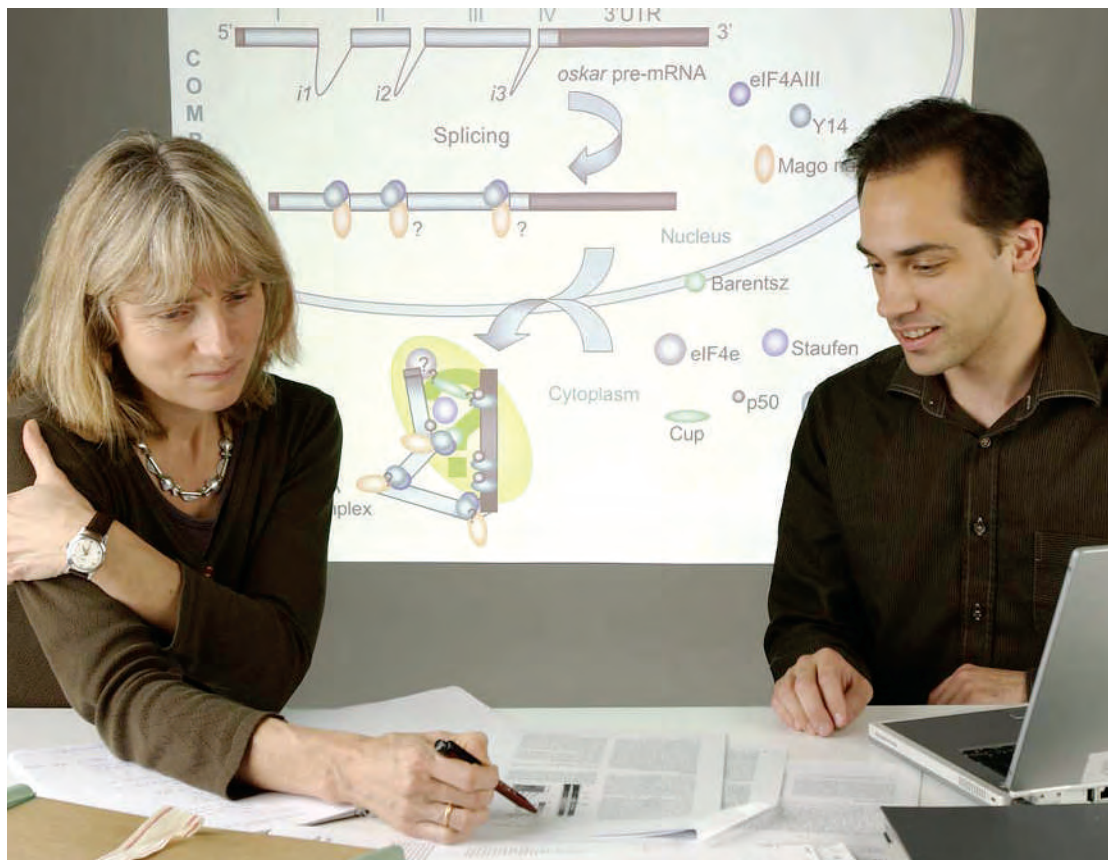
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
 Patrick Phelan
 e-mail: p.phelan@nature.com

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Learning to mentor

Having a good mentor can determine the direction and probability of success for a young researcher. But mentoring takes skill, and institutions are paying attention to their training, says **Virginia Gewin**.



Lucy Godley, an assistant professor of medicine at the University of Chicago, has had mentors all her life. Beginning in high school and through college and graduate study at the Pritzker School of Medicine at the University of Chicago, where she had an appointed mentor, she has been helped by advisers and confidants to navigate the various issues of academic life. But now it is her turn to be a mentor, and she finds the responsibility daunting.

"My first graduate student starts soon and I hope to take the best of what I had and pass it on," she says, expressing excitement but also trepidation.

Godley is not alone. Making the transition from having a mentor to being one is harder than one might think. Managing people, rather than experiments, is unfamiliar territory for many early-career scientists. Given the number of horror stories — about mentors who are uncommunicative, absent or even competitive — it is clear that not all graduate students have positive experiences to draw from. Institutions, individual departments and even online mentoring services are trying to identify the factors that can make more mentoring experiences positive.

Identifying traits

Understanding a student's career aspirations is often the first step towards tailoring a mentoring style to an individual student, say these mentors. Identifying a student's strengths and weaknesses may involve a

"I went in with the naive notion that students would be just like me," says Hopi Hoekstra, left.

difficult conversation, but this will often be one of the most fruitful. To get at a student's individual needs may require a fairly in-depth exchange.

The Federation of American Societies for Experimental Biology has developed an informal checklist, dubbed the individual development plan (IDP), to offer a guide to mentors and students (see 'Basic steps for mentors', left). Such checklists, although new to science programmes, are a staple in the business world.

To promote consistency, some institutions are creating mentoring programmes or putting new emphasis on programmes already in place. Both the Howard Hughes Medical Institute (HHMI)/Burroughs Wellcome Fund and the European Molecular Biology Organization (EMBO) have added mentoring to their programmes for training young investigators in a variety of lab management skills. EMBO's Young Investigator Programme also provides a mentor for one year, including one paid visit to the protégé's workplace.

Hopi Hoekstra, now an assistant professor of biology at the University of California, San Diego, says that the HHMI course helped her avoid mistakes from the beginning. Nonetheless, she says, she still had to learn to adapt her mentoring style to meet individual needs.

"I went in with the naive notion that students would be just like me, and that's absolutely not the case," says

BASIC STEPS FOR MENTORS

- Become familiar with available opportunities
- Discuss opportunities with the postdoc
- Review individual development plan (IDP)
- Establish regular review of progress
- Help revise the IDP as needed

MINORITY MENTORING

Sometimes, female or minority students may have issues beyond those that come up with students in general. There are mentoring organizations that address these specific groups. For example, the Women in Natural Sciences Developing Opportunities in Mentoring programme at the University of Texas assembles both one-to-one mentoring pairs and small groups to help undergraduate, graduate and postdoc students find or become mentors.

Some initiatives, such as MentorSET — a UK-based mentoring programme (sponsored by the Association for Women in Science and Engineering and the Women's Engineering Society) — are designed to help retain women in science, engineering and technology. MentorSET offers free mentoring workshops to attract mentors. The American Biomedical Research Conference for Minority Students holds conferences to encourage minority students to pursue biomedical careers.

Creating an affinity group where scientists are helping students culturally, as well as professionally, is important, says Alberto Roca, a biochemist at the University of California, Irvine, and founder of the Minority Postdoc Summit forum. **V.G.**



Taught to teach: postdocs are increasingly getting training on how to mentor.

Hoekstra. "You can't just have one mentoring style." And that's where more personalized attention can become helpful (see 'Minority mentoring', above).

This may have particular resonance in Europe, where graduate students tend to arrive with different levels of education, experience and cultural expectations, says Anne Ephrussi, associate dean of graduate study at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. EMBL's approach involves a mandatory two-month intensive course to explain basic research and the framework of laboratory research, and to establish a tight network where students meet all group leaders.

"EMBL is the most non-hierarchical a place could be and still remain highly functional," says Ephrussi, explaining the decision to bring more structure into the mentoring process.

Like many other institutions, the Research Institute of Molecular Pathology (IMP) in Vienna, Austria, uses senior faculty members as mentors for junior colleagues. This is not a formal process — junior faculty members typically take the initiative to find a natural match with a senior colleague. But, says Jan Michael Peters, a cellular biologist at the IMP, the physical design of the institution may encourage the relationship as well as more informal networking and advising. Unlike other European laboratories, where groups tend to be more separated, he suggests

that the IMP's layout forces people to bump into each other. The central cafeteria next to the lecture hall provides numerous opportunities to meet people for a coffee and a chat following mandatory attendance at student talks, he says.

Reaching out for more ideas

Communicating and networking, outside the mentoring relationship and beyond a particular academic programme, have been identified as crucial to successful career-building. Sandra Schmid, chair of the department of cell biology at the Scripps Research Institute in La Jolla, California, suggests that mentors should encourage students to talk to other people about their work, and form collaborations to generate more work — all in an effort to establish their value to the scientific community.

Indeed, encouraging students to have several mentors, ideally outside their immediate project, helps build their networking skills. An outside perspective can be invaluable, particularly if one isn't set on a traditional career route. What if a student decides to go into industry and the mentor has only had experience in academia? Pointing students towards industry mentoring resources can help, says biochemist Sonja Lorenz, a PhD student at the University of Oxford, UK, who took part in a Novartis-sponsored mentoring programme to expand her circle of mentors and offer industry career options.

To serve an increasing demand for external mentors in mathematics, science and engineering, the online not-for-profit free service MentorNet, based at San José State University in California, has matched thousands of people with mentors since 1998. It even offers online training tutorials to those willing to serve as a mentor. In addition to case studies, mentors receive regular coaching suggestions. It started out as an industry resource, but a pilot academic programme was set up to meet demand from graduate students and postdocs. The biggest challenge is finding enough tenured faculty members, says Carol Muller, MentorNet's chief executive.

Practical advice may only be part of a mentor's role. Success may mean helping a protégé to find or create an environment that is comfortable to work in.

"I want a place where people want to come to work," says Ashutosh Chilkoti, associate director of the Center for Biologically Inspired Materials and Material Systems at Duke University in Durham, North Carolina. Rather than monitor postdoc hours in the lab, Chilkoti rewards hard work. Although he offers practical career and technical advice — such as how to get a lab running faster and more cheaply, or how to write grant proposals — he believes this advice pales in comparison to what he calls his most important job: conveying a passion for the work. His students seem to agree. Chilkoti was recently voted a top postdoc mentor in a survey by *Science*.

Lucy Godley can attest to the sustaining effect of a mentor's enthusiasm. Confidence born from the encouragement of past mentors helped her through the rough patches of starting her own lab, she says. Becoming an assistant professor turned out to be harder than she expected. "It took all that encouragement for all those years to keep me doing it," she says.

Virginia Gewin is a freelance science writer based in Portland, Oregon.

WEB LINKS

- FASEB guide to mentoring
- ♦ www.faseb.org/opa/ppp/educ/idp.html
- The Howard Hughes Medical Institute Lab Management
- ♦ www.hhmi.org/grants/office/graduate/labmanagement.html
- European Molecular Biology Organization Young Investigator Programme
- ♦ www.embo.org/projects/yip
- Women in (Natural) Sciences Developing Opportunities in Mentoring programme at the University of Texas
- ♦ www.utexas.edu/cons/wins/wisdom
- MentorSET
- ♦ www.mentorset.org.uk
- Minority Postdoc Summit
- ♦ www.minoritypostdoc.org

MOVERS

**Giovanni Galizia, professor of neurobiology,
University of Konstanz, Germany**



2003–05: Associate professor, Department of Entomology, University of California, Riverside

1999–2005: Head, independent research group, Berlin

1995–99: Postdoctoral fellow, Free University Berlin

1993–95: Postdoctoral fellow, Max Planck Institute, Tübingen

1989–93: Doctoral research, University of Cambridge, UK

Giovanni Galizia is exactly the type of researcher of which German science policy-makers dream. He is young, gifted, committed — and about to return to his native Germany after a successful stay abroad. Until recently Galizia combined working in California with leading an independent research group funded by the Volkswagen Foundation and the Free University of Berlin.

Since the 42-year-old researcher started thinking about science, he says, he has been inspired by the big questions: how the world works, what nature is all about, and how its phenomena are connected. He didn't have a preference for one specific subject area initially, although in time he gravitated towards a career in biology and mathematics, and then began to focus on the neurobiology of odour processing in insect brains.

The basic architecture of the olfactory system is similar in most animals, explains Galizia, and he hopes that he will decode the neural connectivity in insects and then be able to extrapolate that knowledge to other species. By combining single-cell analysis with studies of entire cell populations, he intends to understand a model neural network in detail. But he doesn't know how long this will take. "There is still a hard nut to crack, because the necessary technological approaches have not yet been developed," he explains.

Galizia's interest in science extends beyond the lab and into process and policy. As speaker and head of the scientific policy committee at the Young Academy — a German joint academic project to establish the promotion of young scholars — for five years, he promoted issues of concern to young scientists. He plans to draw on his US experience to promote ideas that could improve Germany's university system. He says that US universities look at other systems around the world in order to adopt the best available models and solutions, a practice he believes should be more widely used at German universities.

With career uncertainty and the absence of a tenure-track system, young scientists in Germany are put in an environment that is not optimal for focusing solely on their research, says Galizia. Therefore, he adds, it is essential to create conditions that will make research careers more predictable and attractive to those who have proved their talent for science.

"The widespread feeling here is that whether or not you make it in science is a question of chance," he says. "What's really lacking is a sense of trust in the system." ■

SCIENTISTS & SOCIETIES

A collective approach

Scientists everywhere — in academia as well as industry — face many of the same work issues at some point during their careers. People get laid off, for instance, or decide to switch jobs, or need to take time off for illness or child care.

To face these challenges, Swedish scientists have historically banded together, rather than going it alone. The Swedish Association of Scientists has, for over 50 years, provided researchers with a collective voice. The organization's unity has been in great demand lately, as biotechnology companies merge, and telecommunication and information-technology firms face greater financial pressures.

The organization has 21,000 members with a university degree in the field of science; more than 25% have a PhD and 20% hold a managerial position. The group is part of the Swedish Confederation of Professional Associations, which unites more than 570,000 members from different professional associations.

In Sweden, most terms of employment are regulated in collective agreements, and these tend to be negotiated by unions and professional associations. According to Swedish legislation, the professional associations and traditional trade

unions are the main advocates for employees' rights.

The Swedish Association of Scientists helps to negotiate terms for members when they leave jobs. And when members suffer job losses, as a result of lay-offs or bankruptcy, the association can supplement their income with insurance, as well as help them find training or other job opportunities.

As a professional organization, it also offers members individual career and salary coaching as well as legal advice. Collectively the group pools its knowledge and resources to lobby for better working conditions, including those within the lab, and on quality-of-life issues and medical care.

This has proved to be a very effective way not just to improve working conditions for Sweden's scientists but also to increase efficiency, and therefore productivity, for its researchers, their institutions and for the country as a whole.

Organizing scientists has worked well in Sweden, but it may not be an option in countries that prohibit scientists from unionizing. Perhaps professional organizations can still offer members advocacy and support. ■

Marita Teräs is editor of the Swedish Association of Scientists' newsletter.
♦ www.naturvetareforbundet.se

GRADUATE JOURNAL

A study in time

Graduate school is a time warp. I'm sure of it. This month, I entered my seventh year of graduate school. The past six years of work seemed to move at radically different paces from each other.

The first two years went by fairly quickly. There was a lot to accomplish: making new friends, completing class requirements, doing rotations in four different labs, choosing a lab and finally starting my thesis work. Things were moving along and I was moving right along with everything.

In my third and fourth years, time felt as if it was slowing down and my life dragged. A major project that seemed to be going well imploded, as did a long-term relationship. Every day was endless, the simplest tasks took too long and I accomplished little. There were moments when I felt time was standing still.

In my fifth year, work began to move along again and life accelerated. Friends and classmates started defending and graduating. Meetings with my thesis committee became more frequent and more important. In my sixth year, the pressures to finish a project, to write a paper, to make decisions about postgraduate school employment pushed the speed of time into warp drive.

My head spins when I think how quickly those 12 months passed. I'm nervous that the next 12 will move even faster. There is too much to accomplish, and if time continues to move at this pace, I'm afraid I won't be able to do it all. ■

Anne Margaret Lee is at Harvard University, Boston, Massachusetts.

Don't mention the 'F' word

Raising brows.

Neil Mathur

"But Mr President, funding femtotechnology will lead to untold advances in medicine, information technology and defence." A piercing look expressed the president's scepticism. No more broken promises would be tolerated.

"Admittedly we had a few hiccups with the earlier technologies, but look, it wasn't that bad. Only 10% of lab heads were successfully convicted, and only 5% of the labs consumed themselves."

Mechanically, the president raised a disapproving eyebrow. "And just think of all the advances..."

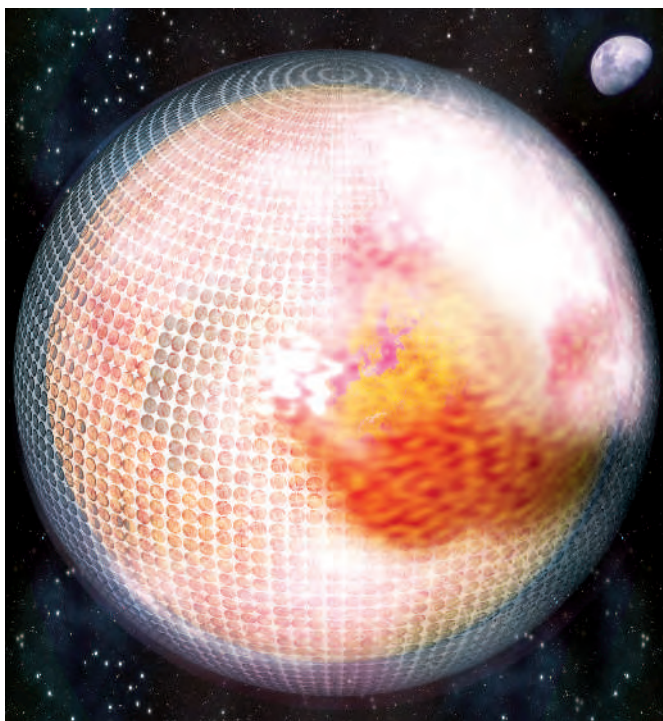
Moving swiftly on, the applicant continued: "As I set out in my agenda, I am going to explain why the previous problems won't plague us this time round."

The president's eyebrow raised itself one notch further.

"Addressing small structures was next to impossible in earlier programmes because everybody was working at disparate length scales. It would be like you trying to address an individual by speaking to a crowd." Another notch. "But the proposed femtotechnology programme will overcome this by cascading the structures from the previous technologies. In fact, it will be just like what you do when you spread the word through the party apparatchiks." The penultimate notch, and still the issues of reproducibility and scale-up to negotiate.

"Reproducibility will be a prerequisite for permanent publication. Yes, I know it is unfortunate that in some cases whole journals had to be reclassified by libraries as science fiction. But this time there will be checks and balances before it gets to the criminal stage. For example, papers will be deleted after two years if they have not been either verified directly or positively cited. And when this happens, all associated honours, promotions, prizes, grants and even centres will be automatically rescinded." Down a notch: so far seven Nobels had been redesignated IgNobels, somewhat spoiling the original IgNobel competition conceived in far more innocent days.

"Scale-up will be routinely achieved by the picobot drones. These can be programmed to implement designs of arbitrary complexity, and the magnitude of the end product is simply determined by the number of drones employed." Back up to the penultimate notch. "Now, wait. There were a lot of advances under the previous programmes and the whole reason that we don't call it 'self-assembly' any more is because we really can do it."



The use of picobot drones would be the unpopular part of the programme to say the least. This is because the 5% of labs that consumed themselves were working on picobot drones. And when the doomed labs had finished consuming themselves, they moved on to the surrounding cities and then just kept on going. This is what precipitated the Evacuation. The doom-sayers had therefore been proved right. Not in every detail, however. For example, the goo to which planet Earth got reduced was not grey, but orange.

A lot of folk were particularly upset about having to relocate to the Moon. Outdoor enthusiasts probably suffered the most. For example, the lunar seas were not much use to the surfing community — most of whom went mad. However, the proponents of small things pleaded mitigation by pointing to the fact that without picobot drones, it would

have been impossible to build the Moon colony. That said, the drones were misbehaving: mutants would often attach themselves to the colonists and draw precious nutrients.

The picobot drones therefore generated by turns, feelings of hatred (for munching through the Earth), gratitude (for building the Moon colony) and irritation (for acting like lunar ticks). Eventually, attitudes towards the drones came to dominate all

conversations. But this is connected with the fact that it was forbidden to mention the picobot clones; or even think about them. In this and other respects, George Orwell — by calling his book 1984 — predicted his vision of totalitarianism only 100 years too soon.

The first picobot clones were built in order to assemble themselves into the cell parts required to make animal muscle tissue. This work was initially funded by a fabulously wealthy restaurateur who wanted interweaving and sculptured cuts of meat for his expensive establishments. Of course, growing meat in this way meant that it was no longer necessary to kill animals. So the project took off on the grounds of animal welfare. Ironically,

all Earth's animals became extinct anyway because there was no ark to take them to the Moon.

The alert reader might have noticed that the evidence linking the picobot drones to the destruction of the Earth was entirely circumstantial. In fact the drones were innocent: it was the clones. Of course the system did not permit the colonists to speculate about this possibility. For if it did, they might have also come to speculate about the clones' ability to differentiate and assemble themselves into, say, the president of the Moon. In fact it was the mention of 'self-assembly' that pushed the leader's eyebrow to its highest level.

"Get out," said the president, "and never come back."

Neil Mathur is in the Department of Materials Science and Metallurgy, University of Cambridge, Cambridge CB2 3QZ, UK.

JACEY